



सत्यमेव जयते

INDIAN AGRICULTURAL
RESEARCH INSTITUTE, New Delhi

I.A.R.I.6.

GIP NLK—H-3 I.A.R.I.—10-5 55—15,000

PROCEEDINGS
OF THE
ROYAL SOCIETY OF LONDON

SERIES A. MATHEMATICAL AND PHYSICAL SCIENCES

VOL 198

LONDON

Published for the Royal Society by the
Cambridge University Press
Bentley House, N.W.1

7 September 1949

Printed in Great Britain at the University Press, Cambridge
(Brooke Crutchley, University Printer)
and published by the Cambridge University Press
Cambridge, and Bentley House, London
Agents for Canada and India: Macmillan

CONTENTS

SERIES A VOLUME 198

No. A 1052. 22 July 1949

	PAGE
The molecular orbital theory of chemical valency. I. The determination of molecular orbitals. By Sir John Lennard-Jones, F.R.S.	1
The molecular orbital theory of chemical valency. II. Equivalent orbitals in molecules of known symmetry. By Sir John Lennard-Jones, F.R.S.	14
A note on polar air-mass modification. By R. Frost	27
Unified field theory in a curvature-free five-dimensional manifold. By J. G. Bennett, R. L. Brown and M. W. Thring	39
The heats of formation of free CN and free CH ₂ , and the relationship between <i>D</i> (CO), <i>D</i> (CN) and <i>D</i> (N ₂). By L. H. Long	62
The behaviour of waves on tidal streams. By N. F. Barber	81
Kinetic theory of diffusion in gases and liquids. I. Diffusion and the Brownian motion. By L. M. Yang	94
Eddy diffusion of water vapour and heat near the ground. By F. Pasquill	116

No. A 1053. 15 August 1949

Thé Royal Greenwich Observatory. By Sir Harold Spencer Jones, F.R.S. (Plates 1 to 4)	141
A note on the Riesz method and the method of residues. By F. C. Auluck and D. S. Kothari	170
Hot-wire investigation of the wake behind cylinders at low Reynolds numbers. By L. S. G. Kovásznay	174
Plastic deformation of silver chloride. I. Internal stresses and the glide mechanism. By J. F. Nye. (Plates 5 to 8)	190
One-dimensional dislocations. I. Static theory. By F. C. Frank and J. H. van der Merwe	205
One-dimensional dislocations. II. Misfitting monolayers and oriented overgrowth. By F. C. Frank and J. H. van der Merwe	216
The rate of evaporation of droplets. II. The influence of changes of temperature and of the surrounding gas on the rate of evaporation of drops of di- <i>n</i> -butyl phthalate. By J. Birks and R. S. Bradley	226
The rate of evaporation of droplets. III. Vapour pressures and rates of evaporation of straight-chain paraffin hydrocarbons. By R. S. Bradley and A. D. Shellard	239
The autoxidation of tetralin. By C. H. Bamford and M. J. S. Dewar	252
The dissociation energy of the N-N bond in hydrazine. By M. Szwarc	267
The dissociation energy of the C-N bond in benzylamine. By M. Szwarc	285

No. A 1054. 22 August 1949

PAGE

The work of the National Institute for Medical Research. By Sir Charles Harington, F.R.S. (Plates 9 and 10)	293
Kinetics of the base-catalyzed halogenation of some ketones and esters. By R. P. Bell, F.R.S., E. Gelles and Eva Möller	308
The catalytic hydrogenation of methyl elaeostearate, and of mixtures of elaeostearic with other polyethenoid long-chain esters. By T. D. Hilditch, F.R.S. and S. P. Pathak	323
Initiation of solid explosives by impact and friction: the influence of grit. By F. P. Bowden, F.R.S. and O. A. Gurton	337
Birth and growth of explosion in liquids and solids initiated by impact and friction. By F. P. Bowden, F.R.S. and O. A. Gurton. (Plates 11 to 14)	350
Influence of entrapped gas on initiation of explosion in liquids and solids. By A. Yoffe. (Plate 15)	373
The effect of diffusion of the main reactants on flame speeds in gases. By J. Corner	388
Paramagnetic resonance in the copper Tutton salts. By B. Bleaney, R. P. Penrose and Betty I. Plumptre	406
The theory of plane plastic strain for anisotropic metals. By R. Hill	428
A theory of the film phenomena of liquid helium II. By H. N. V. Temperley	438

No. A 1055. 7 September 1949

Spontaneous emulsification of pure xylene in an aqueous solution through mere adsorption of a detergent in the interface. By A. Kaminski and McBain, F.R.S. (Plate 16)	447
The diffraction of blast. I. By M. J. Lighthill	454
Kinetic theory of diffusion in gases and liquids. II. General kinetic theory of liquids mixtures. By L. M. Yang	471
The behaviour of continuous stanchions. By J. F. Baker, M. R. Horne and J. W. Roderick	493
Studies in polymerization. V. The polymerization of vinyl acetate. By G. Dixon-Lewis	510
The ignition of solid explosive media by hot wires. By E. Jones	523
Heisenberg's <i>S</i> matrix for a system of many particles. By R. J. Eden	540
Penetration of magnetic field into superconductors. II. Measurements by the Casimir method. By E. Laurmann and D. Shoenberg	560
Electroviscosity IV. Some extensions of the theory of flow of liquids in narrow channels. By G. A. Elton and F. G. Hirschler	581
Index	590

Corrigendum

Proceedings of the Royal Society A, vol. 197, plates 5 and 6

The illustrations for plates 5 and 6 should be transposed, but the legends should remain as printed, i.e. the illustrations described in the legend to plate 5 actually appear on plate 6 and *vice versa*.

The Royal Society regrets the inconvenience caused to the reader.

The molecular orbital theory of chemical valency

I. The determination of molecular orbitals

BY SIR JOHN LENNARD-JONES, F.R.S.

(Received 20 August 1948, publication delayed at request of author)

In the molecular orbital theory of valency the electrons are assigned to the whole molecule rather than to atoms or to other localized parts. While the method has advantages in dealing with the properties of a molecule as a whole, such as its energy states, the extension of each orbital over the molecular framework is a disadvantage when dealing with localized properties such as directed bonds. This paper deals in a general way with the equations which molecular orbitals must satisfy, allowing for the exchange of electrons between orbitals. It is then shown that when molecules have properties of symmetry the equations can be transformed so as to be satisfied by orbitals which have the property of equivalence. These can be regarded under certain conditions as directed orbitals and the conditions for these are discussed. To illustrate the method molecules of the type XY_2 are considered.

1. INTRODUCTION

Modern theories of chemical valency are based on the principles of wave mechanics and the theories have met with success so long as it has been possible to apply these principles satisfactorily to problems of molecular structure. The difficulties and limitations of existing theories arise not so much from any intrinsic deficiency in the principles involved but rather in the intractable nature of the calculations to which they lead. It is thus a central problem in theoretical chemistry to try to simplify wherever possible the methods of calculations and to reduce to their simplest form the wave equations which have to be solved. The object of this and the following paper is to obtain the equations which must be satisfied by the orbitals which electrons occupy in molecules and to try to transform them by the use of the symmetry properties of molecules to a more convenient form.

There are at present two main methods of calculation. One of them, called the electron pair method, is essentially a generalization of the calculation used by Heitler and London for the hydrogen molecule. The essential feature of this treatment is that when the electrons of interacting atoms are paired in spin a molecular state is obtained which is lower in energy than that of the isolated atoms. The development of this method has resulted in a close understanding of the structure of a wide range of molecules and has led to a quantitative treatment, though necessarily an approximate one, of the interactions of atoms in molecules. One advantage of this method is that it is based on various 'states' of a molecule such as can be represented by conventional structural formulae. The results can be interpreted in terms of these basic structures and provide the chemist with a picture of a molecular state as a superposition of acceptable structural formulae. A further advantage of the method is that it permits the use of directed atomic orbitals (Pauling 1931, 1946; Slater 1931) and introduces the concept of directed valencies. The criterion used in preparing an atom for interaction with other atoms is that the wave function should be as concentrated as possible in the direction of union. It is possible to combine

(or hybridize) atomic wave functions such that a number of equivalent orbitals can be regarded as directed in particular directions. Thus the carbon atom can be prepared to have four equivalent orbitals in tetrahedral directions, or alternatively to have three equivalent orbitals in trigonal directions with a fourth orbital in a direction at right angles to the plane of the other three.

These two characteristic features of the electron pair theory have led to its extensive use in the theory of molecular structure and to its general acceptance by chemists as a satisfying picture of the processes at work in determining the properties of molecules.

The other main method of calculating the properties of molecules is called the molecular orbital theory. It was first used for diatomic molecules (Lennard-Jones 1929) and later extended to ethylene and conjugated molecules (Hückel 1930, 1937). It has been extensively used and developed by Mulliken (1932*a*, 1932*b*, 1933, 1941) who has been particularly successful in interpreting the spectroscopic properties of molecules in terms of the individual orbitals to which the electrons in a molecule have to be assigned. Selection rules have been obtained in an elegant way from the symmetry properties of the electronic orbitals.

In this method the electrons are supposed assigned to the molecule as a whole, consisting as it does of nuclei in a certain symmetrical arrangement relative to each other and other electrons in certain orbitals consistent with that symmetry. This method has marked advantages in dealing with the energy states of certain molecules, particularly conjugated molecules, for it determines the sequence of energy levels of the individual electrons in molecules just as in atoms and thus indicates the sequence of closed shells of electrons. In conformity with the Pauli exclusion principle the electrons can be assigned to the orbitals to give the state of lowest energy, or in some cases excited states, and so the energy content of the molecule can be calculated relative to any convenient standard.

Each of the two main methods of calculating the structure of molecules has its characteristic advantages and often both are used, the one to supplement the other. The calculations based on the methods are necessarily approximate at present, but in spite of this the results obtained for molecular energies are in most cases in general agreement.

The molecular orbital method has the merit of being more fundamental in its approach, for it treats electrons as belonging to the whole molecule rather than to atoms or to other localized parts of the molecule. But the extension of each orbital over the molecular framework is a disadvantage when it is necessary to deal with localized properties such as localized bonds. In this and the following paper the foundations of the orbital theory are examined and equations obtained which the molecular orbitals must satisfy when interchanges between orbitals are allowed. It is shown that when molecules possess properties of symmetry these equations can be transformed in such a way that they are satisfied by orbitals which have the property of equivalence. Such equivalent orbitals are identical as regards their distribution and differ only in their orientation; they are, in fact, directed orbitals which can be regarded as forming localized bonds. The method by which these results are obtained is based on the group theory, which is a powerful aid in dealing

with molecules of known symmetry. The equations satisfied by equivalent orbitals are simpler than those satisfied by molecular orbitals in that they involve functions of only one type and occur in sets so that the solution of any one implies a solution of the rest. To illustrate the method a simple example is worked out in this paper, a molecule of type XY_2 being considered, but it is hoped to deal more generally with equivalent orbitals in a later paper and to give the equations which such orbitals must satisfy. Recent advances in computing devices encourage the hope that eventually it may be possible to solve these equations, as has already been done for atoms, and so by an inverse transformation to derive molecular orbitals.

2. MOLECULAR ORBITAL WAVE FUNCTIONS

The wave equation in atomic co-ordinates for a molecule containing N electrons and having nuclei of atomic numbers Z_α, Z_β , is (neglecting spin)

$$\left\{ \sum_{j=1}^N H_j + \sum_{j>k}^N (1/r_{jk}) - E \right\} \Psi = 0, \quad (2.01)$$

where

$$H_j = -(\tfrac{1}{2}) \nabla_j^2 - \sum_{\alpha} Z_{\alpha}/r_{\alpha j}, \quad (2.02)$$

and E is the energy of the electrons, the total energy being

$$E + \sum Z_{\alpha} Z_{\beta} / r_{\alpha \beta}.$$

Z_{α} is the atomic number of a typical nucleus and $r_{\alpha j}$ the distance of the j th electron from it. The wave function Ψ and the energy of the electrons E will, of course, depend on the nuclear configurations determined by the nuclear distances $r_{\alpha \beta}$.

The molecular orbital method of approximating to the solution of the above equation is to suppose that another equation can be found of the type

$$\left\{ \sum_{j=1}^N H_j^* - E \right\} \Psi = 0, \quad (2.03)$$

where

$$H_j^* = -(\tfrac{1}{2}) \nabla_j^2 - v_j, \quad (2.04)$$

and v_j is a function of the j th set of co-ordinates alone. The equation (2.03) can then be solved in terms of functions which satisfy

$$(H_j^* - \epsilon) \psi = 0. \quad (2.05)$$

This is an equation in three dimensional space and its solutions ψ_1, ψ_2, \dots with energies $\epsilon_1, \epsilon_2, \dots$ give distributions in ordinary space which are called molecular orbitals. The solution of the equation (2.03) then consists of products of the molecular orbital wave functions, or sums of such products and the possible energy values consist of sums of the energies $\epsilon_1, \epsilon_2, \dots$ of individual orbitals.

With each molecular orbital we may associate one or other of the wave functions in the spin co-ordinates, denoted by $\alpha(\sigma)$ and $\beta(\sigma)$, the co-ordinate σ denoting spin and taking the two values $\pm \frac{1}{2}$. The spin functions may be taken to be orthogonal and normalized in the usual way.

The wave function of the N electron system, which satisfies (2.03), is then, including spin, of the type

$$\Phi = \det \{ \psi_1(1) \alpha(1) \psi_2(2) \beta(2) \dots \}, \quad (2.06)$$

and the energy is the sum of the energy contributions of the individual orbitals corresponding to ψ_1, ψ_2 , which are assumed occupied by electrons. Since according to the exclusion principle not more than two electrons occupy each orbital, the energy is given by

$$E = \sum_r c_r \epsilon_r, \quad (2.07)$$

where c_r is equal to 0, 1, or 2 according to the number of electrons in the orbital of energy ϵ_r . In this summation it has been supposed that all the orbitals have different energies, but if some of them should have coincident energies as in cases of degeneracy, then the above expression must be generalized accordingly, the permissible values of c_r being those given above for each member of a degenerate set.

There will usually be many solutions of the type (2.06) for the same energy (2.07) owing to the possibility of assigning electrons to different orbitals with the same energy.

One of the great advantages of this method is that the electrons are conceived as moving in potentials fields v_j which depend on the position of each electron alone, the effect of the other electrons on the one considered being assumed averaged over their appropriate distributions. The field v_j is assumed to be similar to that of the nuclear framework and so is unchanged when equivalent nuclei are supposed interchanged by any symmetry operation. The powerful methods of the group theory then provide much information about the symmetry properties of the orbitals and so of the composite wave functions (2.06). These symmetry properties permit a determination of the possible changes from one molecular state to another and so give the selection rules for the absorption and emission of light. The energies of the orbitals for fields v_j of different symmetries appropriate to a wide range of molecules have been worked out. They usually involve integrals involving v_j and so cannot be determined unless the field is known, but the difficulty can often be side-stepped by regarding the integrals as parameters which can be determined from one observable property and then used to evaluate others.

3. A MORE GENERAL TREATMENT OF THE ORBITAL THEORY

The disadvantage of the method just outlined is that it does not make clear what is the relation between the equation (2.03) and the actual one (2.01) which is to be solved. A similar difficulty arose in dealing with the structure of atoms. At first electrons were dealt with one at a time and the method of self-consistent fields was devised and used extensively by Hartree (1927, 1930, 1947). Later an improvement was introduced by Fock (1930) who used the variation method to transform equation (2.01) involving $3N$ co-ordinates to N equations each in three co-ordinates. This treatment can be applied equally well to molecules. The method to be used in this paper is that given by the author (1931) in the theory of atoms containing many electrons.

It is assumed that the solution of the wave equation of a molecule can be expressed in terms of functions of the co-ordinates of individual electrons (that is, without the introduction of the relative co-ordinates of electrons). The most general way of expressing the wave functions of all the electrons is by sums of products of individual wave functions and, in particular, to get the right properties of antisymmetry in the co-ordinates (Pauli principle) by a determinant of the type (2.06). When there are many such determinants corresponding to the same total orbital energy, then it may be necessary to consider sums of these. For the ground state of a molecule, or for any state which corresponds to a set of complete electron shells, one determinant is sufficient. It is also possible to deal by a similar treatment with an electronic configuration of a molecule corresponding to a number of closed shells and a number of electrons outside these shells in different orbitals with the same spin. Such a configuration corresponds to a term of multiplicity $(2S+1)$, where S is the number of outer electrons.

To cover this latter case we suppose that there are p orbitals in the closed shells, each occupied by an electron of α and β spin, and $(q-p)$ outer electrons each with a spin α (or β). We then have altogether q electrons with α spin and p with β . The determinant is then

$$\Phi = \det \{ \psi_1(1) \alpha(\sigma_1) \psi_2(2) \alpha(\sigma_2) \dots \psi_q(q) \alpha(\sigma_q) \psi_1(q+1) \beta(\sigma_{q+1}) \dots \psi_p(N) \beta(\sigma_N) \}, \quad (3.01)$$

We have so far made no assumptions about the ψ 's, the orbital wave functions, but it is possible to derive from the last p functions a set which are orthogonal to each other. Such a transformation will not alter Φ except by a numerical factor. Similar linear combinations can be made of the first p functions containing α instead of β spin factors. Finally linear combinations of the functions so derived and the remaining $(q-p)$ functions can be made so that all the functions used in Φ are orthogonal, either because of their spatial properties or because of their spin factors. Normalization of each of the component functions will only alter Φ by a numerical factor and so we may suppose this adjustment made without loss of generality.

The Euler variation equation corresponding to the wave equation (2.01) is

$$\int \delta \Phi \left\{ \sum_j H_j + \sum_{j>k} (1/r_{jk}) - E \right\} \Phi d\tau = 0 \quad (3.02)$$

$$\text{or} \quad \int \left(\sum_i \delta_i \Phi \right) \left\{ \sum_j H_j + \sum_{j>k} (1/r_{jk}) - E \right\} \Phi d\tau = 0, \quad (3.03)$$

where δ_i indicates a variation of functions containing the i th co-ordinates wherever they appear in Φ .

$$\text{Now} \quad \bar{\Phi} \Phi = \rho = \det \{ \rho(i, j) \}, \quad (3.04)$$

$$\text{where} \quad \rho(i, j) = \sum_{l=1}^N \bar{\phi}_l(i) \phi_l(j) \quad (3.05)$$

and the functions ϕ_l include an orbital factor ψ and a spin factor. There are N such functions, the first q containing an α spin factor and the remainder a β -spin factor. The determinant in (3.04) contains N rows and columns, i and j separately taking all values from 1 to N .

Equation (3.03) can be written in this notation as

$$\int \left(\sum_i \delta_i \right) \left(\sum_j H_j + \sum_{j>k} (1/r_{jk}) \right) \rho d\tau = E \int \left(\sum_i \delta_i \right) \rho d\tau, \quad (3.06)$$

where δ_i now indicates a variation of the functions $\bar{\phi}$ whenever they contain the i th co-ordinates in ρ . Thus

$$\delta_i \rho(i, j) = \sum_{l=1}^N \delta_i \bar{\phi}_l(i) \phi_l(j). \quad (3.07)$$

Similarly H_j operates on the ϕ functions in ρ which contain the j th co-ordinates, so that

$$H_j \rho(i, j) = \sum_{l=1}^N \bar{\phi}_l(i) H_j \phi_l(j), \quad (3.08)$$

and

$$\delta_i H_j \rho(i, j) = \sum_{l=1}^N \delta_i \bar{\phi}_l(i) H_j \phi_l(j). \quad (3.09)$$

Thus $\delta_i H_j \rho$ represents a determinant in which terms such as those of (3.07) and (3.08) occur in the i th row and the j th column, the common member being (3.09). If the terms occurring in this determinant are denoted by ω_{ij} , where i indicates the row and j the column, then

$$\int \omega_{ik} \omega_{kj} d\tau_k = \omega_{ij} \quad (3.10)$$

for all values of i and j including the i th row and the j th column. It is thus possible to integrate $(\delta_i H_j) \rho$ over all variables except the i th and j th and the result is

$$\begin{aligned} \int (\delta_i H_j) \rho d\tau &= (N-2)! \int \delta_i H_j \begin{vmatrix} \rho(i, i) & \rho(i, j) \\ \rho(j, i) & \rho(j, j) \end{vmatrix} d\tau_i d\tau_j \\ &= (N-2)! \left\{ \left(\sum_m H_{mm} \right) \sum_n \int \delta \bar{\psi}_n \psi_n dx - S \sum_{m,n} H_{mn} \int \delta \bar{\psi}_n \psi_m dx \right\}, \end{aligned} \quad (3.11)$$

where in the last equation the integration over the spin co-ordinates has been carried out and the symbol S represents a summation over those pairs of functions, ψ_m, ψ_n , which have the same spin factor in Φ ; those with different spin factors are orthogonal and vanish on integration. The space co-ordinates for the remaining integrals are denoted by x , a suffix no longer being necessary. In the above equation

$$H_{mn} = \int \bar{\psi}_m H \psi_n dx. \quad (3.12)$$

In a similar way it is found that

$$\int \delta_i H_i \rho d\tau = (N-1)! \sum_n \int \delta \bar{\psi}_n H \psi_n dx, \quad (3.13)$$

$$\int \left(\sum_i \delta_i \right) \sum_{j \neq i} (1/r_{ij}) \rho d\tau = N! \left\{ \sum_n \int \delta \bar{\psi}_n \psi_n V(x) dx - S \sum_{m,n} \int \delta \bar{\psi}_n \psi_m G_{mn}(x) dx \right\}, \quad (3.14)$$

where

$$G_{mn}(x) = \int \bar{\psi}_m(x') \psi_n(x') (1/r) dx' \quad (3.15)$$

is the potential of a distribution $\bar{\psi}_m \psi_n$ at a point whose co-ordinates are represented by x . $V(x)$ is the potential due to all the distributions $\bar{\psi}_m \psi_m$, viz.

$$V(x) = \sum_n G_{nn}(x). \quad (3.16)$$

Similarly $(\delta_i/r_{jk})\rho$ can readily be integrated over all variables except the i th, j th and k th and ρ is reduced to a three column determinant in these variables. The remaining integrations then lead to the equation

$$\begin{aligned} \int \left(\sum_i \delta_i \right) \left(\sum_{j>k} (1/r_{jk}) \rho d\tau = \frac{1}{2}(N!) \left\{ \sum_{l,m} (lm | G | lm) - S \sum_{l,m} (lm | G | ml) \right\} \sum_n \int \delta \bar{\psi}_n \psi_n dx \right. \\ \left. - \frac{1}{2}(N!) \left\{ \sum_{l,m,n} S (lm | G | ln) - S \sum_{l,m,n} (lm | G | nl) \right\} \int \delta \bar{\psi}_n \psi_m dx, \right. \end{aligned} \quad (3.17)$$

$$\text{where} \quad (lm | G | nl) = \int \bar{\psi}_l(j) \bar{\psi}_m(k) \psi_n(j) \psi_i(k) (1/r_{jk}) dx_j dx_k \quad (3.18)$$

and similar integrals for the other terms. The summation S is over those sets of l, m, n three ψ functions which have the same spin in Φ . The energy E is given by

$$\begin{aligned} E &= \int \left(\sum_j H_j + \sum_{j>k} (1/r_{jk}) \right) \rho d\tau \div \int \rho d\tau \\ &= \sum_n H_{nn} + \frac{1}{2} \left\{ \sum_{l,m} (lm | G | lm) - S \sum_{l,m} (lm | G | ml) \right\}. \end{aligned} \quad (3.19)$$

When the above results are substituted in equation (3.06), we obtain an integral containing the differentials $\delta \bar{\psi}_n$ each multiplied by a factor. If equation (3.06) is to be satisfied for all variations $\delta \bar{\psi}_n$, the factors must separately vanish, and the result is a set of linear differential equations in three co-ordinates instead of Schrödinger's equation in $3N$ co-ordinates. These equations are

$$\begin{aligned} [H + V(x)] \epsilon_n \psi_n(x) - S_m^{(n)} \left\{ G_{mn}(x) + H_{mn} + \sum_l (lm | G | ln) \right\} \psi_m \\ + S_{l,m}^{(n)} (lm | G | nl) \psi_m = 0, \end{aligned} \quad (3.20)$$

where ϵ_n is the number of times ψ_n occurs in Φ and is either 1 or 2; $S_m^{(n)}$ and $S_{l,m}^{(n)}$ are summations over those functions ψ_m and ψ_l which have the same spin factor as ψ_n in Φ . For a molecule containing only closed shells with every orbital occupied by an electron with α spin and one with β spin, we have ϵ_n equal to 2 and

$$S_m^{(n)} = 2 \sum_1^{\frac{1}{2}N}, \quad S_{l,m}^{(n)} = 2 \sum_1^{\frac{1}{2}N} \sum_1^{\frac{1}{2}N}.$$

The set of equations (3.20), being equivalent to the wave equation (2.01), is accurate subject to the assumption of a solution of the type (3.01).

4. MOLECULES OF KNOWN SYMMETRY

In attempting to solve the equations (3.25) we are to remember that the nuclear framework will in many molecules of interest have certain properties of symmetry. This is implied in the properties of the Hamiltonian H , which is given by

$$H = -(\frac{1}{2}) \nabla^2 - \sum_{\alpha} Z_{\alpha}/r_{\alpha}, \quad (4.01)$$

where r_{α} is the distance of a point from the nucleus of atomic charge Z_{α} .

If we confine the present discussion to a molecule in its normal state consisting of closed shells, equation (3.20) becomes

$$\{H + V'_n(x) - E_{nn}\} \psi_n = \sum'_m \{G_{mn}(x) + E_{mn}\} \psi_m, \quad (4.02)$$

$$\text{where} \quad E_{mn} = H_{mn} + \sum_l \langle lm | G | ln \rangle - \sum_l \langle lm | G | nl \rangle \quad (4.03)$$

$$\text{and} \quad V'_n(x) = \sum^*_m G_{mm}(x) \quad (4.04)$$

and \sum'_m used in (4.02) is a summation over all space orbitals, counted once only, omitting the orbital ψ_n , while \sum^* is a summation over all space orbitals, counted twice, except ψ_n which occurs only once.

The equations (4.02) can also be written in the abbreviated form

$$(\mathfrak{H})(\psi) = 0, \quad (4.05)$$

where (\mathfrak{H}) is an operator matrix, the diagonal elements being

$$\mathfrak{H}_{nn} = H + V'_n(x) - E_{nn} \quad (4.06)$$

$$\text{and the remainder being} \quad \mathfrak{H}_{mn} = -G_{mn}(x) - E_{mn}; \quad (4.07)$$

(ψ) denotes the array of occupied ψ functions.

If equation (4.02) is multiplied on the left by $\overline{\psi}_m$ and integrated over the whole of space, the result given in (4.03) is obtained. E_{nn} is clearly the energy of an electron in the orbital ψ_n moving in the presence of the field of the nuclei (H_{nn}) and the field of the other electrons, averaged over their respective charge distributions $\left(\sum'_m \langle mn | G_{mn} | \rangle \right)$ with an additional contribution $\left(- \sum'_m \langle mn | G | nm \rangle \right)$ owing to the field $G_{mn}(x)$, which arises from the possibility of interchange between one orbital and another.

The total energy of the electrons is not obtained simply by adding the energies of the individual orbitals because the electrostatic contributions, represented by the last two terms in E_{nn} , would be counted twice; it is given by equation (3.19).

A good approximation to the solution of the equations (4.02) for atoms has been obtained by neglecting the right-hand side of the equation and consequently also the non-diagonal matrix elements $\langle ln | G | nl \rangle$ which occur in E_{nn} owing to the presence of $G_{mn}(x)$. In justification for this step it may be observed that whereas all the members $G_{mm}(x)$ occurring in $V'_n(x)$ are positive in all parts of space, the functions $G_{mn}(x)$ are positive in some regions of space and negative in others; in

fact $G_{mn}(x)$ is the electrostatic potential at any point due to a unit charge distributed over space according to $\bar{\psi}_m\psi_n$ and the latter function, when integrated over the whole of space, vanishes because of the orthogonal property. For a similar reason H_{mn} the only remaining factor of ψ_m on the right-hand side is also neglected.

We may adopt a similar procedure for molecules and attempt as a first approximation to find solutions of

$$\{H + V'_n(x) - \mathcal{E}_n\} \psi_n = 0 \quad (4.08)$$

with

$$\mathcal{E}_n = H_{nn} + \sum_l' (ln | G | ln). \quad (4.09)$$

The methods of group theory can be applied to equation (4.05) to find the type of solution ψ_n for a field H of given symmetry. It is one of the results of this theory that the electron distribution of closed shells has the same symmetry as that of the nuclear framework. Thus all distributions $\bar{\psi}_m\psi_m$ for non-degenerate orbitals ψ_m have this symmetry and also $\sum_k \bar{\psi}_k\psi_k$ when ψ_k belongs to a degenerate set. It

follows that $V'_n(x)$ in (4.05) has the same symmetry as H for all orbitals ψ_m which are non-degenerate. When ψ_n belongs to a degenerate set, $V'_n(x)$ may not have the same symmetry properties, but in this case we may suppose the rest of the functions $\sum_k' \bar{\psi}_k\psi_k$ in the same set so averaged to produce the right symmetry. This

procedure was necessary in many electron atoms when an electron in a p or d orbital was being considered. We shall accordingly suppose that the operator in equation (4.05) has the same symmetry properties as H in all cases.

If, for example, the potential field has the symmetry C_{2v} like a triangular molecule of the type XY_2 , which has an axis of symmetry (z), a plane of symmetry through it (yz) at right angles to the plane of the nuclei and a plane of symmetry (xz) through the nuclei, the possible types of orbital conform to the scheme in table 1. There are four types of orbital and their behaviour, when rotated through 180° about the axis of $z(C_2)$, or reflected in the xz plane (σ_v), or reflected in the yz plane (σ'_v), is given by the numbers in the table. All the orbitals in this example are single, and so $V'_n(x)$ always has the symmetry of the group.

TABLE 1. TYPES OF MOLECULAR ORBITAL IN A FIELD OF SYMMETRY C_{2v}

	E	C_2	σ_v	σ'_v
A_1	1	1	1	1
A_2	1	1	-1	-1
B_1	1	-1	1	-1
B_2	1	-1	-1	1

We may thus imagine a set of solutions of equation (4.08) worked out conforming to the appropriate symmetry requirements such as are given in table 1. Such solutions of different symmetries will be orthogonal to each other and so will satisfy the conditions laid down in §3 above. The determinantal wave function defined in equation (3.01), must be built up of orbital wave functions satisfying the right symmetry requirements as laid down by the group theory.

Once solutions of this kind have been worked out, the more accurate equations (4.02) can be solved by a process of successive approximation. Thus the appropriate

functions ψ_m can be substituted on the right-hand side and $G_{mn}(x)$ and E_{mn} evaluated, and then the right-hand side regarded as a perturbation of the ψ_n which satisfies the left-hand side alone.

Though the functions ψ_m substituted on the right-hand side may have symmetry properties which are different from ψ_n , we note that $G_{mn}(x)\psi_m$ has the same symmetry as ψ_n , if ψ_m belongs to a non-degenerate set. If ψ_m belongs to a degenerate set, then $\sum_m G_{mn}(x)\psi_m$ summed over the members of the set has the same symmetry as ψ_n . Further, if ψ_m and ψ_n have different symmetry properties, then E_{mn} vanishes. Hence the right-hand side of (4.02) has the same symmetry as ψ_n and only those numbers E_{mn} remain for which m and n have the same symmetry.

Thus a set of functions may be derived which solve the equations (4.02) in the sense that they give self-consistency and each function has the same symmetry properties as the approximate solution derived from (4.08). Each of the orbitals so determined will have an energy

$$E_{nn} = H_{nn} + \sum_l (ln | G | ln) - \sum_l (ln | G | nl) \quad (4.10)$$

and the total energy of the system will be given by equation (3.19).

5. DIRECTED ORBITALS

The functions determined by the above method give distributions which extend throughout the nuclear framework and are thus molecular orbitals in the sense in which that term is widely used. For spectroscopic purposes molecular orbitals have advantages in that the energy levels of the individual electrons may be supposed known so that electron transitions to other excited states may be understood and described, while the selection rules may be easily deduced from the symmetry properties of the orbitals. From the point of view of valency the method has disadvantages in that the chemist is accustomed to think of electrons in chemical bonds as localized. Also the bonds are known from experiment to be orientated in definite directions. This feature has been well reproduced by the method of Pauling (1931) and by Slater (1931) of taking linear combinations of atomic orbitals and obtaining orbitals with directional properties. Thus an s and three p orbitals can be combined to give an orthogonal set of four equivalent orbitals directed towards the vertices of a regular tetrahedron (cf. Hultgren 1932; Kimball 1940 for further examples). An illuminating discussion of the particular case of the methane molecule has been given by Van Vleck (1933), assuming that molecular orbitals can be expressed as linear sums of atomic orbitals. The treatment has recently been taken further in an interesting way for this and other molecules by Coulson (1949). It would be an advantage if directed orbitals could be obtained in terms of molecular orbitals and shown to be part of the same logical scheme. It is the object of this and the following paper to deal with this problem.

We consider in the first place a simple example of a molecule such as XY_2 of C_{2v} symmetry and inquire whether two equivalent orbitals can be derived, each associated with a XY bond. Let the two orbitals be denoted by χ_1 and χ_2 . Then under the operations of the group, as given in table 1, we find that χ_1 and χ_2 change as in table 2.

TABLE 2. THE BEHAVIOUR OF EQUIVALENT ORBITALS
IN A FIELD OF SYMMETRY C_{2v}

C_{2v}	E	C_2	σ_v	σ'_v
χ_1	χ_1	χ_2	χ_1	χ_2
χ_2	χ_2	χ_1	χ_2	χ_1
(χ_1, χ_2)	2	0	2	0

The character scheme of the set χ_1, χ_2 is given in the last line. From this it follows by comparison with table 1 that the set must be made up of orbitals of types A_1 and B_1 . We suppose therefore, that there are two molecular orbitals, which we label as ψ_1 and ψ_2 , one being of symmetry A_1 and the other of B_1 . For simplicity we will consider these orbitals alone apart from any others which may be occupied in the molecule, and suppose that each contains two electrons, one of α -spin and the other of β -spin. The equations satisfied by ψ_1 and ψ_2 are obtained from (4.02), viz.

$$\left. \begin{aligned} \{H + V'_1(x) - E_1\} \psi_1 &= G_{21}(x) \psi_2, \\ \{H + V'_2(x) - E_2\} \psi_2 &= G_{12}(x) \psi_1, \end{aligned} \right\} \quad (5.01)$$

$V'_1(x)$ being the electrostatic field on an electron in orbital ψ_1 and similarly for $V'_2(x)$. E_1 and E_2 are here written for E_{11} and E_{22} .

Now χ_1 and χ_2 are to be linear sums of ψ_1 and ψ_2 and are to be orthogonal. If ψ_1 and ψ_2 are also supposed normalized, then χ_1 and χ_2 satisfy these conditions and are also normalized, if

$$\chi_1 = (1/\sqrt{2})(\psi_1 + \psi_2), \quad \chi_2 = (1/\sqrt{2})(\psi_1 - \psi_2). \quad (5.02)$$

It follows that χ_1 and χ_2 can equally well be used in the molecular wave function Φ and the same methods used to determine the equations satisfied by χ_1 and χ_2 . Thus from equation (4.02) we have at once

$$\{H + v'_1(x) - e_{11}\} \chi_1 = \{g_{21}(x) + e_{21}\} \chi_2, \quad (5.03)$$

$$\{g_{12}(x) + e_{12}\} \chi_1 = \{H + v'_2(x) - e_{22}\} \chi_2, \quad (5.04)$$

where $v'_1(x)$ contains terms such as

$$g_{11}(x) = \int \chi_1(x') \chi_1(x') (1/r) dx' \quad (5.05)$$

and e_{12} is the energy term derived from (4.03) using the χ functions. Thus

$$e_{12} = h_{12} + \sum_l (l1 | g | l2) - \sum_l (l1 | g | 2l), \quad (5.06)$$

where h_{12} is the matrix element of H with respect to the χ functions and $(\kappa\lambda | g | \mu\nu)$ similarly means an integral of $(1/r_{12})$ over the appropriate χ functions.

The equations (5.01) for the ψ functions are similar in form to those above, except that E_{12} vanishes because the two ψ functions have different symmetry properties. On the other hand we find that

$$h_{11} = \frac{1}{2}(H_{11} + H_{22}) = h_{22}, \quad (5.07)$$

$$h_{12} = \frac{1}{2}(H_{11} - H_{22}) \quad (5.08)$$

(and $H_{12} = 0$).

The electrostatic contribution to the energy terms is invariant under the transformation and

$$(12 | g | 12) - (12 | g | 21) = (12 | G | 12) - (12 | G | 21),$$

so that we have
$$e_{11} = e_{22} = \frac{1}{2}(E_1 + E_2), \quad (5.10)$$

whereas
$$e_{12} = e_{21} = \frac{1}{2}(E_1 - E_2). \quad (5.11)$$

The equations for the χ 's therefore become

$$\left. \begin{aligned} \{H + v'_1(x) - \frac{1}{2}(E_1 + E_2)\} \chi_1 &= \{g_{21}(x) + \frac{1}{2}(E_1 - E_2)\} \chi_2, \\ \{H + v'_2(x) - \frac{1}{2}(E_1 + E_2)\} \chi_2 &= \{g_{12}(x) + \frac{1}{2}(E_1 - E_2)\} \chi_1. \end{aligned} \right\} \quad (5.12)$$

The same result can, of course, be obtained by a direct transformation of the equations (5.01), using the relations

$$\left. \begin{aligned} g_{11}(x) &= \frac{1}{2}(G_{11} + G_{22}) + G_{12}, \\ g_{22}(x) &= \frac{1}{2}(G_{11} + G_{22}) - G_{12}, \\ g_{12}(x) &= g_{21}(x) = \frac{1}{2}(G_{11} - G_{22}). \end{aligned} \right\} \quad (5.13)$$

The interesting feature of equations (5.12) is that they are identical except for notation. χ_2 is similar to χ_1 for it is its mirror image in the yz plane of symmetry of the molecule. It is to be noted also that the function $g_{12}(x)$ involves the product of χ_1 and χ_2 in the integrand, and when there is little overlap, it will be small. If further the orbitals ψ_1 and ψ_2 have nearly the same energies, $(E_1 - E_2)$ will be small, so that the whole of the right-hand side of (5.12) will be small. Hence to a first approximation we may write

$$\{H + v'_1(x) - E\} \chi_1 = 0 \quad (5.14)$$

and the problem is reduced to finding a localized orbital (doubly occupied) in one bond only XY of the molecule XY_2 subject to the electrostatic effect of the other Y nucleus and a similar distribution in the other XY bond. Since the latter bond will partly screen the other Y nucleus, the effect of the second bond may perhaps be represented by a simple electrostatic distribution such as a dipole. The localized orbital determined by (5.14) may thus be regarded as a molecular orbital embracing two nuclei and perturbed by the presence of a similar orbital in the second bond.

In order that the χ directed orbitals may differ appreciably from the ψ molecular orbitals, it is necessary that ψ_1 and ψ_2 should overlap to a substantial amount. If, for example, ψ_1 and ψ_2 were states which had energies differing widely from each other, then the nodal surfaces of one would be more concentrated near the nuclei than the other. In the outer parts of the molecule there would be little difference between χ_1 and χ_2 . The terms $g_{12}(x)$ and $\frac{1}{2}(E_1 - E_2)$ occurring on the right-hand side of equation (5.12) would then both be large and the approximation equation (5.14) would no longer be valid. Hence we see that the condition for directed orbitals is that the energy difference $(E_1 - E_2)$ between molecular orbitals must be small.

To obtain directed orbitals from molecular orbitals all that has been done is to apply certain mathematical transformations; the molecule remains exactly as before. The molecule can thus be described equally well by molecular orbitals or

by directed orbitals. From the wave theory point of view either is equally valid. It is rather a question of what particular property of the molecule we are interested in. If it is the energy levels and the differences in energy due to excitation, then the molecular orbital method gives the answer. We are not then concerned so much with the location of the electrons and they may be dispersed throughout the molecule. But when it is a question of the electron distribution in bonds, then a superposition of orbitals to give directed properties becomes appropriate. The energy of individual electrons is then of less importance and a knowledge of distribution implies less precise information about energy. But if an electron is excited, it is not possible to say from which (directed) bond it came. It may have come from any one of a set of equivalent bonds which is another way of saying that it has been excited from a molecular orbital.

REFERENCES

- Coulson, C. A. 1949 *J. Chim. Phys.* (in the Press—the author is indebted for an advance copy of the manuscript.)
 Fock, V. 1930 *Z. Phys.* **61**, 126.
 Hartree, D. R. 1928 *Proc. Camb. Phil. Soc.* **24**, 89, 111.
 Hartree, D. R. 1930 *Proc. Camb. Phil. Soc.* **25**, 225.
 Hartree, D. R. 1947 For full references see *Reports on Progress in Physics*, **11**, 113.
 Hückel, E. 1930 *Z. Phys.* **60**, 423.
 Hückel, E. 1937 *Z. Elektrochem.* **43**, 752, 827.
 Hultgren, R. 1932 *Phys. Rev.* **40**, 891.
 Kimball, G. E. 1940 *J. Chem. Phys.* **8**, 188.
 Lennard-Jones, J. E. 1929 *Trans. Faraday Soc.* **25**, 668.
 Lennard-Jones, J. E. 1931 *Proc. Camb. Phil. Soc.* **27**, 469.
 Mulliken, R. S. 1932*a* *Phys. Rev.* **40**, 55.
 Mulliken, R. S. 1932*b* *Phys. Rev.* **41**, 49, 751.
 Mulliken, R. S. 1933 *Phys. Rev.* **43**, 279.
 Mulliken, R. S. 1941 For further references see *Reports on Progress in Physics*, **8**, 231.
 Pauling, L. 1931 *J. Amer. Chem. Soc.* **53**, 1367.
 Pauling, L. 1946 *The Nature of the Chemical Bond*. Cornell University Press.
 Slater, J. C. 1931 *Phys. Rev.* **37**, 481.
 Van Vleck, J. H. 1933 *J. Chem. Phys.* **1**, 177, 219.

The molecular orbital theory of chemical valency

II. Equivalent orbitals in molecules of known symmetry

By SIR JOHN LENNARD-JONES, F.R.S.

(Received 17 December 1948)

The paper aims at providing a connecting link between the theory of molecular orbitals and the theory of localized bonds. An examination is made of the fundamental equations which must be satisfied by molecular orbitals in fields of known symmetry, particularly in molecules of the type XY_n . It is shown by the methods of the group theory that these equations can be transformed to others which involve sets of equivalent functions. These are associated with *equivalent orbitals* which have the property of being identical as regards distribution in space and differ only in their orientation. It is shown that under certain conditions these can be regarded as localized orbitals associated with particular bonds. General formulae are obtained and applied to the particular cases of molecules of trigonal, tetrahedral and octahedral symmetry.

1. INTRODUCTION

The object of the preceding paper (Lennard-Jones 1949) was to find the relation between the electron-pair theory of valency, which deals with the interaction of electrons in localized bonds, and the molecular orbital theory of valency, which treats each electron as moving in the field of the whole molecular framework. In order to make progress with either of these theories, it is customary to assume, as a starting point, that each electron can be assigned to an orbital, which can be described by a function of three spatial co-ordinates. In the one case these orbitals are assumed to be localized between an atom and an interacting neighbour, while in the other the orbitals are distributed throughout the molecule, representing in fact stationary states of an electron in the field of all the nuclei and the averaged field of the other electrons. In each method the wave function of the whole molecule is expressed as a combination of functions of the individual electrons. To conform to Pauli's principle, these one-electron wave functions are combined in the form of a determinant.

In the preceding paper equations were obtained for the molecular orbitals to which the electrons can be assigned in a molecule in its normal, unexcited state. It was then shown that in a symmetrical molecule such as XY_2 these equations could be transformed so as to be satisfied by a distribution about one of the bonds XY and a similar distribution about the other XY bond. These distributions were called equivalent orbitals, because each was the mirror image of the other in the plane of symmetry of the molecule. In this paper the results are generalized to apply to a molecule of the type XY_n , in which there is a central atom X and n other atoms or groups distributed in some symmetrical way about it.

2. EQUATIONS SATISFIED BY MOLECULAR ORBITALS

Equations were given in part I (equation (3.20)) to determine the wave functions of the molecular orbitals of a molecule either in its ground state or in an excited state, provided that this state arises from closed shells and outer electrons possessing the

same spin. While the substance of the present paper will be concerned mainly with molecules in their ground states, or more accurately with those which consist of a set of closed shells of paired electrons, it may be of interest to consider the form of the equations for the more general case.

When there are more electrons with one kind of spin than another the equations for molecular orbitals are unsymmetrical, and it is necessary to be careful as to the precise meaning to be attached to each term and each summation. An alternative way of writing the equations (3.20) of part I is as follows:

$$[H + V(x)](\epsilon_n^\alpha + \epsilon_n^\beta) \psi_n(x) - \sum_m (\epsilon_m^\alpha \epsilon_n^\alpha + \epsilon_m^\beta \epsilon_n^\beta) [G_{mn}(x) + H_{mn} + J_{mn}^\alpha + J_{mn}^\beta] \psi_m + \sum_m (\epsilon_m^\alpha \epsilon_n^\alpha K_{mn}^\alpha + \epsilon_m^\beta \epsilon_n^\beta K_{mn}^\beta) \psi_m = 0. \quad (2.01)$$

$$\text{In these equations} \quad \epsilon_n^\alpha = 1 \text{ or } 0; \quad \epsilon_n^\beta = 1 \text{ or } 0, \quad (2.02)$$

according as ψ_n is associated with an α spin or not; and similarly with a β spin or not. H is the Schrödinger operator for a single electron moving in the field of the nuclei of the molecule and H_{mn} its matrix element relative to ψ_n and ψ_m . The definition of the various operators is as follows:

$$G_{mn}(x_j) = \int \bar{\psi}_m(x_i) \psi_n(x_i) (1/r_{ij}) dx_i, \quad (2.03)$$

where x_i and x_j stand for the spatial co-ordinates of two electrons,

$$V(x) = \sum_i (\epsilon_i^\alpha + \epsilon_i^\beta) G_{ii}(x), \quad (2.04)$$

$$J_{mn}^\alpha = \sum_i \epsilon_i^\alpha (lm | G | ln), \quad (2.05)$$

$$K_{mn}^\alpha = \sum_i \epsilon_i^\alpha (lm | G | nl), \quad (2.06)$$

$$\text{with} \quad (\kappa\lambda | G | \mu\nu) = \int \bar{\psi}_\kappa(x_i) \bar{\psi}_\lambda(x_j) (1/r_{ij}) \psi_\mu(x_i) \psi_\nu(x_j) dx_i dx_j. \quad (2.07)$$

We note, in fact, that J_{mn}^α is the average value of $G_{mn}(x)$ when taken over all orbitals occupied by electrons with α spin, viz.

$$J_{mn}^\alpha = \sum_i \epsilon_i^\alpha \int \bar{\psi}_i(x_j) G_{mn}(x_j) \psi_i(x_j) dx_j. \quad (2.08)$$

There are similar definitions of J_{mn}^β , K_{mn}^α and K_{mn}^β as sums over orbitals containing electrons with α or β spins.

If we suppose that there are p orbitals which are occupied by electrons with both α and β spins, and that there are q outer electrons with one kind of spin only, say α spin, then it is possible to express the equations (2.01), of which there are $(p+q)$ altogether, in the form

$$(\mathcal{H})(\psi) = 0, \quad (2.09)$$

where (\mathcal{H}) is an array of operators in matrix form and (ψ) represents the set of $(p+q)$ orbitals, expressed as a single-column matrix. The matrix (\mathcal{H}) is of $(p+q)$ dimensions and takes the form

$$\begin{pmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{pmatrix}$$

where \mathcal{A} is of p dimensions and \mathcal{D} of q dimensions.

If we write $V'_{nn}(x) = V(x) - G_{nn}(x)$, $J_{mn} = J_{mn}^{\alpha} + J_{mn}^{\beta}$,

we can express the diagonal elements as

$$\mathcal{H}_{nn} = (\epsilon_n^{\alpha} + \epsilon_n^{\beta}) (H + V'_{nn}(x) - H_{nn} - J_{nn}) + \epsilon_n^{\alpha} K_{nn}^{\alpha} + \epsilon_n^{\beta} K_{nn}^{\beta}, \quad (2.10)$$

and the non-diagonal elements as

$$\mathcal{H}_{mn} = -(\epsilon_m^{\alpha} \epsilon_n^{\alpha} + \epsilon_m^{\beta} \epsilon_n^{\beta}) (G_{mn}(x) + H_{mn} + J_{mn}) + \epsilon_m^{\alpha} \epsilon_n^{\alpha} K_{mn}^{\alpha} + \epsilon_m^{\beta} \epsilon_n^{\beta} K_{mn}^{\beta}. \quad (2.11)$$

The equations take a simpler and more symmetrical form when all the electrons are paired, for then the matrix becomes of type \mathcal{A} only. In this case we can write

$$\mathcal{H}_{nn} = 2(H + V'_{nn}(x) - E_{nn}), \quad (2.12)$$

$$\mathcal{H}_{mn} = -2(G_{mn}(x) + E_{mn}), \quad (2.13)$$

where

$$E_{mn} = H_{mn} + J_{mn} - K_{mn}^{\alpha}, \quad (2.14)$$

$$\text{so that the equations are } \mathcal{H}_{nn} \psi_n + \sum'_m \mathcal{H}_{mn} \psi_m = 0. \quad (2.15)$$

We note that E_{mn} is the energy of an electron in a stationary state in the field of the nuclei and the space-charge distribution of the electrons in all other orbitals, less a contribution due to the field $G_{mn}(x)$, which arises from the possibility of interchange of electrons between orbitals.

We have already discussed in part I methods of solving these equations. All that need be said here is that if the exchange property of the electrons is neglected, the equations reduce to

$$\mathcal{H}_{nn} \psi_n = 0, \quad (2.16)$$

and solutions of this equation correspond to stationary states in the field of the nuclei and the electrostatic field of the electrons in all other occupied orbitals, represented by the term $V'_{nn}(x)$. The electrostatic field due to a set of closed shells will have the symmetry of the nuclear framework, and so for non-degenerate states the appropriate wave function will belong to one of the irreducible representations of the symmetry group. There is a difficulty in the case of degenerate states, for then $V'_{nn}(x)$ includes not only the field of a set of closed shells but also the field of the remaining electrons in the same shell. The latter field will not, in general, have the same symmetry as the nuclear framework. To preserve the symmetry properties of the wave functions in degenerate sets it will be necessary to suppose the field on each electron in the set so averaged that each moves in the *same* field, and, moreover, that this field has the full symmetry of the molecule. The functions so obtained from equation (2.16) will then be self-consistent, not only in the Hartree sense, but also as regards their symmetry properties, for these will then conform to those of the permissible irreducible representations. If these solutions are used as first approximations of equation (2.16), nothing in the subsequent calculations will upset the symmetry properties, though the values of the energy parameters and the wave functions will be changed.

3. SETS OF EQUIVALENT ORBITALS

The equations (2.01) were obtained from a determinantal wave function of the form

$$\Phi = \text{Det} \{ \psi_1(1) \alpha(1) \psi_1(2) \beta(2) \dots \psi_p(2p-1) \alpha(2p-1) \psi_p(2p) \beta(2p) \\ \times \psi_{p+1}(2p+1) \alpha(2p+1) \dots \psi_{p+q}(2p+q) \alpha(2p+q) \}, \quad (3.01)$$

and the properties of the system will not be altered by any transformation which leaves this wave function unchanged. Thus any orthonorm transformation of the functions ψ_1 to ψ_n , which constitute its elements, will not change Φ . It is unchanged when the nuclear framework is subject to any of its symmetry operators, for this is equivalent to multiplying the various columns of the determinant by numerical factors and taking linear sums. The determinant is unaltered except for a numerical factor, and for unitary transformations this factor is unity.

Though such transformations do not alter the system intrinsically, they can be used to simplify the method of solution of the equations (2.01). Thus we endeavour to find for a molecule of known symmetry a set of orbitals which are equivalent in the sense that they are interchangeable under the operations of the group. It is possible to determine from the character table appropriate to any type of molecular symmetry which types of molecular orbitals can be superimposed to produce a set of equivalent orbitals. The method is to find the character system of the equivalent orbitals under all the operations of the group and then to compare the result with the character system of the molecular orbitals (the irreducible representations of the group). In general it is found that a set of equivalent orbitals for a molecule of the type XY_n can be expressed as a sum of a molecular orbital which has the perfect symmetry of the group (the identity representation, usually denoted by Γ_1) and others of single or degenerate type, which behave differently under the operations of the group.

Suppose that E, A, B, \dots, P represent operations of the symmetry group to which a molecule of type XY_n belongs, E being as usual the identity operator. Then the irreducible representations of the group are given by a set of matrices of the kind shown in table 1. Each entry such as (a_j) represents a matrix of dimensions equal to the degeneracy of the representation Γ_j , and (e_j) represents a unit matrix of the same dimensions. If Γ_1 is the identity representation, all the entries in the first row are unity.

TABLE 1. THE REPRESENTATIONS OF A SYMMETRY GROUP OF TYPE XY_n

	E	A	B	\dots	P
Γ_1	(e_1)	(a_1)	(b_1)	\dots	(p_1)
Γ_2	(e_2)	(a_2)	(b_2)	\dots	(p_2)
.....					
Γ_j	(e_j)	(a_j)	(b_j)	\dots	(p_j)
$\chi\text{-set}$	(e_n)	(a_n)	(b_n)	\dots	(π_n)

If a set of equivalent orbitals $\chi_1, \chi_2, \dots, \chi_n$ exists for a molecule XY_n , the effect of the operations E, A, B, \dots, P of the group will be to interchange the orbitals

in various ways. Hence the set will be represented by a series of matrices of the type $(\epsilon_n), (\alpha_n), \dots, (\pi_n)$ shown in the last row of the table. Each will be of dimensions n , and each will contain only unity and zeros, unity occurring once in each row and once in each column.

From a comparison of the character system of the χ -set with that of the $\Gamma_1, \dots, \Gamma_j$ representations, it is possible, if a set of equivalent orbitals exists, to show that the same set of characters can be obtained by a superposition of some of the Γ 's. Thus we may write

$$(\chi) \equiv \sum_j c_j \Gamma_j, \quad (3.02)$$

where c_j is some numerical factor, usually unity or zero. Since the dimensions of the matrix representation of the χ -set is n , the dimensions of the sum of the right-hand side must also be n . The interpretation of this result is that a set of equivalent orbitals can be found by superposition of molecular orbitals, whose symmetry properties are the same as the representations Γ_j included in the above formula.

The relation (3.02) has been used to indicate the types of directed *atomic* orbitals which can be obtained from a superposition of s, p, d, \dots , functions (Kimball 1940). It is equally valid for *molecular* orbitals and it is now our purpose to show that by its use a transformation of the equations satisfied by molecular orbitals can be carried out.

Though the characters of (χ) and $\sum_j c_j \Gamma_j$ are the same, the matrices of the latter representation, when written out in full in their n -dimensional form, will not necessarily be the same as the matrices $(\epsilon_n), (\alpha_n), (\beta_n), \dots, (\pi_n)$ representing the χ -set. Let the matrices of $\sum_j c_j \Gamma_j$ be denoted by $(E_n), (A_n), (B_n), \dots, (P_n)$. It follows from the group theory that there must be a relation between corresponding matrices of the form

$$(A_n) = T(\alpha_n) T^{-1}, \quad (B_n) = T(\beta_n) T^{-1}, \quad \dots, \quad (3.03)$$

where T is some transformation operator; it will, of course, be represented by a matrix of n dimensions, thus

$$T = (t).$$

Further, if (ψ) denotes a set of molecular orbitals which has the representation $\sum_j c_j \Gamma_j$, then the relation between the χ 's and the ψ 's is given by

$$(\chi) = T(\psi). \quad (3.04)$$

Using this transformation the equations to determine the equivalent orbitals (χ) can be obtained from those satisfied by the molecular orbitals (ψ) given in (2.01). Thus the appropriate equations are

$$\mathfrak{T}(\mathcal{H})(\psi) = 0, \quad (3.05)$$

where \mathfrak{T} indicates that every ψ function is subject to a transformation T both where it appears in the matrix elements of (\mathcal{H}) and where it appears in (ψ) ; or in other words, every ψ is changed to a χ .

In order to illustrate the process it will be sufficient to consider the form taken by (3.05) for one set of equivalent orbitals and to limit the discussion to a set of orbitals which are doubly occupied. Equation (3.05) then becomes

$$(\mathcal{H}')(\chi) = 0, \quad (3.06)$$

where (\mathcal{H}') is an operator matrix with

$$\mathcal{H}'_{nn} = 2(H + v'_{nn}(x) - e_{nn}), \quad (3.07)$$

$$\mathcal{H}'_{mn} = -2(g_{mn}(x) + e_{mn}), \quad (3.08)$$

and the functions v'_{nn} , g_{mn} are obtained from $V'_{nn}(x)$, $G_{mn}(x)$ appearing in (2.03) and (2.04) by changing each ψ to one of the χ functions.

To complete the transformation it is necessary to calculate the value of e_{nn} and e_{mn} in terms of quantities which depend on the molecular orbitals. It is found possible to express them in terms of the quantities E_{nn} and E_{mn} defined in equation (2.14). Thus e_{mn} is the quantity derived from E_{mn} when in the various integrals from which it is calculated the equivalent orbital wave functions are substituted for molecular orbital wave functions and the relation (3.04) is used. For the case when the occupied molecular orbitals form closed shells, we have from (2.14)

$$e_{mn} = \mathfrak{L}(E_{mn}) = \mathfrak{L}(H_{mn} + J_{mn} - K_{mn}^{\alpha}). \quad (3.09)$$

Now H_{mn} and J_{mn} are the matrix elements with respect to $\bar{\psi}_m(x_i)$ and $\psi_n(x_i)$ of H and the function $\rho(j, j) (1/r_{ij})$ respectively, where

$$\rho(jj) = \rho^{\alpha}(jj) + \rho^{\beta}(jj) \quad (3.10)$$

and

$$\rho^{\alpha}(jj) = \sum_i \epsilon_i^{\alpha} \bar{\psi}_i(x_j) \psi_i(x_j). \quad (3.11)$$

Similarly

$$K_{mn}^{\alpha} = \int \bar{\psi}_m(x_i) [\rho^{\alpha}(j, i) (1/r_{ij})] \psi_n(x_j) dx_i dx_j, \quad (3.12)$$

where

$$\rho^{\alpha}(j, i) = \sum_t \epsilon_t^{\alpha} \bar{\psi}_t(x_j) \psi_t(x_i). \quad (3.13)$$

Now the summation in the last equation covers a set of molecular orbitals which include all the members of certain degenerate irreducible representations. In such a case $\rho^{\alpha}(j, j)$ and $\rho^{\alpha}(j, i)$ are invariant under any unitary transformation T .

It follows that

$$k_{mn}^{\alpha} = \mathfrak{L}(K_{mn}^{\alpha}) = \int \left[\sum_{\mu} \bar{t}_{m\mu} \bar{\psi}_{\mu}(x_i) \right] \rho^{\alpha}(j, i) (1/r_{ij}) \left[\sum_{\nu} t_{n\nu} \psi_{\nu}(x_i) \right] dx_i = \sum_{\mu, \nu} (\bar{t}_{m\mu} t_{n\nu}) K_{\mu\nu}^{\alpha}, \quad (3.14)$$

with a similar result for h_{mn} and j_{mn} . Hence

$$e_{mn} = \sum_{\mu, \nu} (\bar{t}_{m\mu} t_{n\nu}) E_{\mu\nu}, \quad (3.15)$$

where $t_{m\mu}$ and $t_{n\nu}$ are elements of the matrix (t) which represents the substitution T , and $\bar{t}_{m\mu}$ the conjugate elements when the matrix elements are complex.

This expression will usually simplify since the matrix elements $H_{\mu\nu}$, $J_{\mu\nu}$ and $K_{\mu\nu}$ vanish whenever the functions $\bar{\psi}_{\mu}$ and ψ_{ν} have different symmetries, that is, belong to different irreducible representations of the symmetry group. A similar result holds

when $\bar{\psi}_\mu$ and ψ_ν are different members of a degenerate set, provided they are suitably chosen. Accordingly, for a set of molecular orbitals which belong to different representations, equation (3.15) simplifies to

$$e_{mn} = \sum_{\mu} (\bar{t}_{m\mu} t'_{n\mu}) E_{\mu\mu}. \quad (3.16)$$

With this expression for e_{mn} , or the more general one (3.15), the equations to determine the set of equivalent orbitals are

$$(H + v'_{nn}(x) - e_{nn}) \chi_n = \sum'_m (g_{mn}(x) + e_{mn}) \chi_m, \quad (3.17)$$

and whereas the equations for the molecular orbitals (2.15) involve a set of different functions ψ_n, ψ_m, \dots , these equations involve only one unknown function, for χ_m differs from χ_n only in its orientation; in fact, the equations included in the set (3.17) can be obtained from one another by cyclical changes of the suffixes, or what amounts to the same thing, by supposing the molecule rotated so that the various χ -functions are interchanged. The set of equations is, therefore, equivalent to one equation and the solution of this implies a solution of all. The problem is to find a distribution χ_n , which, together with all the other similar but differently orientated distributions χ_m , renders the equation (3.17) self-consistent.

Once a set of equivalent orbitals has been found to solve the equation with appropriate values of the parameters e_{nn} and e_{mn} , the appropriate molecular orbitals can be derived by using the substitution which is inverse to equation (3.04), viz.

$$(\psi) = T^{-1}(\chi), \quad (3.18)$$

and the energy levels can be similarly derived from a relation which is the inverse of (3.15) and of the type

$$E_{\mu\nu} = \sum_{m,n} (t_{m\mu} \bar{t}_{n\nu}) e_{mn}. \quad (3.19)$$

4. SPECIAL CASES

It is instructive to find the form taken by the above equations in molecules of particular symmetry types. Molecules of C_{2v} symmetry, of which triangular ones like XY_2 are examples, have already been considered in part I. Other types of special interest because of their wide occurrence in chemistry are those of trigonal, tetrahedral and octahedral symmetry.

(i) *Molecules XY_3 of trigonal symmetry (C_{3v})*

For molecules of the C_{3v} type there are two symmetry operators of a rotation of 120° clockwise or anticlockwise about the axis of symmetry and these are denoted by $2C_3$. There are also three planes of symmetry through the axis of symmetry ($3\sigma_v$). There are three types of molecular orbital; two of them (A_1 and A_2) are single and the third (E) is doubly degenerate. The character scheme is shown in table 2. (For a description of the notation used in this section reference may be made to Mulliken (1933) or to Eyring, Walter & Kimball (1946).)

The substitution matrices of three equivalent orbitals, χ_1, χ_2 and χ_3 , under the operations of the group will be respectively of the types

$$(\chi_1 \chi_2 \chi_3): \begin{matrix} E & C_3 & \sigma_v \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{matrix}. \quad (4.01)$$

These matrices may be denoted by $(P)_\chi$, where P stands for any of the operators of the group. The character system is thus that given in the last line of table 2. It follows that the equivalent orbitals can be obtained by a superposition of one A_1 orbital and two degenerate orbitals of type E .

TABLE 2. TYPES OF MOLECULAR ORBITAL IN A FIELD OF SYMMETRY C_{3v}

	E	$2C_3$	$3\sigma_v$
A_1	1	1	1
A_2	1	1	-1
E	2	-1	0
χ	3	0	1

In order to find the matrix of the transformation T it is necessary to examine the matrices of operations of the group in the representation $(A_1 + E)$. Two typical matrices are

$$(A_1 + E): \begin{matrix} C_3 & \sigma_v \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & \sqrt{\frac{3}{4}} \\ 0 & -\sqrt{\frac{3}{4}} & -\frac{1}{2} \end{pmatrix} & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \end{matrix}, \quad (4.02)$$

and the others, apart from E which is a diagonal matrix, are obtained from them by writing the matrices of C_3^2 , $\sigma_v C_3$ and $\sigma_v C_3^2$. These matrices may be denoted by $(P)_{A_1+E}$. It is then found that the substitution T which gives equivalence between corresponding matrices, in the form

$$(P)_\chi = (t) (P)_{A_1+E} (t)^{-1}, \quad (4.03)$$

is

$$(t) = \begin{pmatrix} \sqrt{\frac{1}{3}} & \sqrt{\frac{2}{3}} & 0 \\ \sqrt{\frac{1}{3}} & -\sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{3}} & -\sqrt{\frac{1}{6}} & \sqrt{\frac{1}{2}} \end{pmatrix}. \quad (4.04)$$

The equations for the equivalent orbitals are then given by equations (3.17) with the following values for e_{mn} :

$$e_{11} = e_{22} = e_{33} = \frac{1}{3}(E_{11} + 2E_{22}), \quad (4.05)$$

$$e_{12} = e_{13} = e_{23} = \frac{1}{3}(E_{11} - E_{22}), \quad (4.06)$$

where E_{11} and E_{22} are the energy parameters which would appear in the equations (2.15) for the molecular orbitals A_1 and E respectively;* the energy parameters E_{12} vanish because of the different symmetry of A_1 and E .

* The use of the symbol E to denote an orbital type and an energy E is unfortunate, but this is the notation generally accepted. The energy parameters in this paper always have distinguishing suffixes.

(ii) *Molecules XY_4 of tetrahedral symmetry (T_d)*

In a similar way it is found that four equivalent orbitals in tetrahedral symmetry are obtained from a superposition of $A_1 + T_2$, where A_1 is an orbital with complete tetrahedral symmetry and T_2 is a triply degenerate set. The matrix for the substitution T is given by

$$(t) = \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad (4.07)$$

and the energy parameters e_{mn} by

$$e_{11} = e_{22} = e_{33} = e_{44} = \frac{1}{4}(E_{11} + 3E_{22}), \quad (4.08)$$

$$e_{12} = e_{13} = e_{14} = e_{23} = e_{24} = e_{34} = \frac{1}{4}(E_{11} - E_{22}), \quad (4.09)$$

where E_{11} and E_{22} are the energy parameters associated respectively with the orbitals A_1 and T_2 in equations (2.15). As in the previous example, E_{12} and similar energy parameters all vanish because of the different symmetries of A_1 and T_2 . There are thus four similar equations for the equivalent orbitals χ_1, χ_2, χ_3 and χ_4 , of which the following is typical:

$$(H + v'_{11}(x) - e_{11}) \chi_1 = g_{21}(x) \chi_2 + g_{31}(x) \chi_3 + g_{41}(x) \chi_4 + e_{12}(\chi_2 + \chi_3 + \chi_4). \quad (4.10)$$

If the energy parameters E_{11} and E_{22} associated with A_1 and T_2 should be equal, indicating accidental degeneracy, then the last term in the equation would vanish. This condition was assumed to be necessary in the original treatment of directed tetrahedral orbitals by Pauling (Pauling 1931; Slater 1931; for a discussion of this assumption, see Van Vleck & Sherman, 1935). It is seen from the equation not to be an essential condition for equivalent orbitals, but it is probable that when this condition holds there may be greater concentration of the χ orbitals about their respective axes. These axes, about which the orbitals will have trigonal symmetry, coincide with the respective XY bonds. When the orbitals are concentrated about these axes and fall away rapidly in density elsewhere then the value of χ_1 will be small on the axes of χ_2, χ_3 and χ_4 , at any rate in the neighbourhood of the Y atoms and beyond. The left-hand side of equation (4.10) will then be small for such regions. Under such conditions the $g(x)$ functions will also be small because they involve integrals of the products of different χ -functions. On the other hand, χ_2, χ_3 and χ_4 will have appreciable values on their own axes. Hence consistency in these statements about concentration can only be obtained when e_{12} is small. Presumably the converse holds, viz. that when e_{12} is large the orbitals are not concentrated locally about an individual axis but are spread through the whole molecule.

Under suitable conditions an approximate solution of the equations (4.10) may be obtained by neglecting the right-hand side. The potential energy terms in $v'_{nn}(x)$ represents the electrostatic effects of χ_2, χ_3 and χ_4 on χ_1 and imply a strong screening of all the Y atoms except one. The χ_1 orbital is then effectively a bicentric one subject to the nucleus of X suitably screened, and to one of the Y 's and to the electrostatic action of the other three bonds.

(iii) Molecules XY_6 of octahedral symmetry (O_h)

For molecules of octahedral symmetry (O_h) there are the 24 elements of symmetry of the group O together with a centre of symmetry, and so with the process of inversion there are 48 symmetry operators. There are ten types of molecular orbital, five of them being even in the process of inversion and five uneven. They are usually denoted by A_1, A_2, E, T_1 and T_2 with the suffixes g and u according as they are even or uneven.

From a comparison of the character system of six equivalent orbitals and that of the molecular orbitals of the O_h group, it is found that

$$(\chi) = A_{1g} + E_g + T_{1u},$$

A_{1g} , as usual, being the orbital with the character system unity, E_g being a doubly degenerate orbital and T_{1u} a triply degenerate one.

The substitution matrix is

$$(t) = \begin{pmatrix} \sqrt{\frac{1}{6}} & 0 & -\sqrt{\frac{1}{6}} & 0 & 0 & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{6}} & 0 & -\sqrt{\frac{1}{6}} & 0 & 0 & -\sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{6}} & \frac{1}{2} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{2}} & 0 & 0 \\ \sqrt{\frac{1}{6}} & \frac{1}{2} & \sqrt{\frac{1}{12}} & -\sqrt{\frac{1}{2}} & 0 & 0 \\ \sqrt{\frac{1}{6}} & -\frac{1}{2} & \sqrt{\frac{1}{12}} & 0 & \sqrt{\frac{1}{2}} & 0 \\ \sqrt{\frac{1}{6}} & -\frac{1}{2} & \sqrt{\frac{1}{12}} & 0 & -\sqrt{\frac{1}{2}} & 0 \end{pmatrix}, \quad (4.11)$$

the first column giving the coefficient of A_{1g} , the next two those of E_g , and the last three those of the T_{1u} orbitals. If χ_1 , given by the coefficients of the first row, is directed along an axis labelled x , then χ_2 is directed along $-x$, χ_3 along y , χ_4 along $-y$, χ_5 along z and χ_6 along $-z$.

Using these matrix elements in equation (3.15) and remembering that $E_{\mu\nu}$ vanishes when the suffixes μ and ν refer to orbitals of different symmetry or to orbitals which are different members of a degenerate set, we find that e_{nn} has the same value for all values of n (from 1 to 6), viz.

$$e_{nn} = \frac{1}{6}(E_{11} + 2E_{22} + 3E_{33}), \quad (4.12)$$

where E_{11} , E_{22} and E_{33} refer respectively to the energy parameters of the orbitals A_{1g} , E_g and T_{1u} . Similarly, there are three pairs of equal energy parameters of the type

$$e_{12} = e_{21} = \frac{1}{6}(E_{11} + 2E_{22} - 3E_{33}), \quad (4.13)$$

and twelve pairs of the type

$$e_{13} = e_{31} = \frac{1}{6}(E_{11} - E_{22}). \quad (4.14)$$

The energy parameters given by (4.13) refer to pairs of orbitals which are directed in opposite directions. The rest refer to pairs of orbitals whose axes are perpendicular to each other.

The equations to determine the six equivalent orbitals are then all similar to

$$(H + v'_{11}(x) - e_{11})\chi_1 = \sum_{m=1}^5 g_{m1}(x)\chi_m + e_{12}\chi_2 + e_{13}(\chi_3 + \chi_4 + \chi_5 + \chi_6). \quad (4.15)$$

The interpretation of this equation is clear. The term $v'_{11}(x)$ is the potential energy of an electron in the presence of the electrostatic distribution of the other equivalent orbitals; four of the functions g_{m1} are similar except for orientation and represent

the potential energy due to an electron distribution given by a product of χ_1 and χ_m ; the energy coefficients are equal for the four equivalent orbitals at right angles to χ_1 , as would be expected.

We note, in particular, that both e_{12} and e_{13} vanish whenever $E_{11} = E_{22} = E_{33}$. An example of this occurs in the theory of directed atomic orbitals where for a hydrogen-like atom octahedral orbitals can be obtained by superposition of (s), (p)³ and (d)² orbitals, all of which have the same energy for the same total quantum number.

An interesting property of equivalent orbitals may be inferred from the energy coefficients e_{mn} . Whenever group theory indicates that equivalent orbitals may be obtained from the superposition of *two* types of molecular orbitals, then there are only two energy levels E_{11} and E_{22} , and accordingly only two coefficients e_{11} and e_{12} . This implies that the orbitals are symmetrically related to each other in the sense that the members of every pair are similarly related to each other. Thus for molecules of trigonal symmetry in a plane, three equivalent orbitals are obtained and the angle between each pair is the same, viz. 120°. For tetrahedral symmetry four equivalent orbitals are so related that the angle between any pair is the same. But when, on the other hand, equivalent orbitals are made up of *three* types of molecular orbital, as in the example of octahedral symmetry, the equivalent orbitals are not similarly related to each other. There are then three energy parameters E_{11} , E_{22} and E_{33} and accordingly three parameters e_{11} , e_{12} and e_{13} . Any one such orbital is then related to the rest in two (and only two) ways. Thus in octahedral symmetry each orbital is similarly related to four of the others and then differently related to the sixth. A similar result is obtained for the case of four equivalent orbitals in a planar molecule XY_4 , where the bonds are directed along the axes of a square.

5. THE ELECTRON DISTRIBUTION IN SYMMETRICAL MOLECULES

When the electronic configuration of a molecule can be described by a set of molecular orbitals determined by equations (2.15) and by a determinantal function Φ , given by equation (3.01), the electronic distribution can be obtained by suitable integrations of $\Phi\Phi$, as was shown by Dirac (1931) to be possible for atoms. The expression $\Phi\Phi d\tau_1 \dots d\tau_N$ is the probability of a given configuration of N electrons, that is, the probability that an electron will be found in a given element of volume $d\tau_1$ with prescribed spin, another in $d\tau_2$, and so on. The probability of finding $(N-1)$ electrons in a prescribed configuration irrespective of the position of the N th is obtained by integrating over the variables of the N th electron. The probability of finding two electrons simultaneously in prescribed elements of volume is obtained by integrating over $(N-2)$ electrons. As we are not interested in the particular electrons which appear in the given elements of volume, it is necessary to integrate over the remaining variables in such a way that each configuration $d\tau_3 \dots d\tau_N$ appears only once. The two-electron distribution function is accordingly proportional to

$$\begin{matrix} \sigma(1, 1) & \sigma(1, 2) \\ \sigma(2, 1) & \sigma(2, 2) \end{matrix} d\tau_1 d\tau_2, \quad (5.01)$$

where

$$\sigma(jk) = \rho^\alpha(jk) \bar{\alpha}_j \alpha_k + \rho^\beta(jk) \bar{\beta}_j \beta_k, \quad (5.02)$$

and the ρ -functions are the summations over the wave functions of the occupied orbitals in ordinary spatial co-ordinates, as defined in equations (3.13) and (3.11).

It is of some interest to examine the above distribution in two special cases. In one case all the orbitals are singly occupied with electrons possessing the same spin, and in the other they are doubly occupied with electrons having paired spins. Both of these cases are included in the following treatment.

When the integration is carried out over the spin co-ordinates, the distribution function becomes

$$\sum_i (\epsilon_\alpha + \epsilon_\beta) \bar{\psi}_i(1) \psi_i(1) \sum_m (\epsilon_\alpha + \epsilon_\beta) \bar{\psi}_m(2) \psi_m(2) - \sum_{i,m} (\epsilon_\alpha^2 + \epsilon_\beta^2) \bar{\psi}_i(1) \psi_i(2) \bar{\psi}_m(2) \psi_m(1), \quad (5.03)$$

the last summation implying that contributions arise only from orbitals having the same spin. When the orbitals are singly occupied with the same spin, the appropriate distribution function is obtained by putting $\epsilon_\alpha = 1, \epsilon_\beta = 0$, and when the orbitals are doubly occupied, it is necessary to put $\epsilon_\alpha = 1, \epsilon_\beta = 1$. When this is reduced by cancelling all possible terms, the expression becomes

$$2\epsilon_\alpha \epsilon_\beta \sum_i \bar{\psi}_i(1) \psi_i(1) \bar{\psi}_i(2) \psi_i(2) + (\epsilon_\alpha + \epsilon_\beta)^2 \sum'_{i,m} \bar{\psi}_i(1) \psi_i(1) \bar{\psi}_m(2) \psi_m(2) - (\epsilon_\alpha^2 + \epsilon_\beta^2) \sum'_{i,m} \bar{\psi}_i(1) \psi_i(2) \bar{\psi}_m(2) \psi_m(1). \quad (5.04)$$

The first term, involving \sum_i , arises from paired electrons in the same orbital, the second and third terms, involving \sum' , from electrons in different orbitals.

Now from the earlier sections of this paper it is clear that it is immaterial whether the wave functions used in the above formulae refer to molecular orbitals or to equivalent orbitals. Thus for singly-occupied orbitals with the same spin, the probability distribution is

$$\left\{ \sum'_{\lambda,\mu} \bar{\chi}_\lambda(1) \chi_\lambda(1) \bar{\chi}_\mu(2) \chi_\mu(2) - \sum'_{\lambda,\mu} \bar{\chi}_\lambda(1) \chi_\lambda(2) \bar{\chi}_\mu(2) \chi_\mu(1) \right\} dv_1 dv_2. \quad (5.05)$$

This gives the probability of finding an electron in an element of volume dv_1 and another in an element of volume dv_2 .

In certain molecules we may expect the orbitals χ_λ, χ_μ to be localized in different bonds. If there is pronounced localization, the value of any one of the χ 's may have a well-defined maximum, and about this maximum there may be a region of space within which χ is appreciable, and outside of which χ has negligible value. In such a case let the region of appreciable χ_λ be denoted by ω_λ . The region ω_λ may or may not overlap another region ω_μ . If it does not, the distribution function (5.05) lends itself to a simple physical interpretation.

Let dv_1 be in the region of ω_λ . Then the function (5.05) is vanishingly small unless dv_2 lies in one of the other regions ω_μ , different from ω_λ . Moreover, it is evident that under these conditions the main contribution to the probability distribution (5.05) comes from the first term. The electrostatic interaction of the electrons, which is given by the integration of e^2/r_{12} multiplied by the probability distribution, in that case arises mainly from those configurations in which two electrons are in different

ω -regions and implies repulsion. In such circumstances the molecule would behave as though the ω -regions repelled each other.

For molecules which contain pairs of electrons in sets of orbitals and consist of complete shells, the corresponding probability function is

$$2 \sum_{\lambda} \bar{\chi}_{\lambda}(1) \chi_{\lambda}(1) \bar{\chi}_{\lambda}(2) \chi_{\lambda}(2) + 4 \sum'_{\lambda, \mu} \bar{\chi}_{\lambda}(1) \chi_{\lambda}(1) \bar{\chi}_{\mu}(2) \chi_{\mu}(2) \\ - 2 \sum'_{\lambda, \mu} \bar{\chi}_{\lambda}(1) \chi_{\lambda}(2) \bar{\chi}_{\mu}(2) \chi_{\mu}(1). \quad (5.06)$$

The first term arises from paired electrons in the same orbital and the other terms from electrons in different orbitals. This formula can be interpreted by means of the model described above. For orbitals localized in well-defined ω -regions, this function has its maximum values for elements of volume dv_1 and dv_2 which are either in the same region ω_{λ} (first term) or in different regions ω_{λ} and ω_{μ} (second term). The third term is always small unless regions ω_{λ} and ω_{μ} overlap. The main contributions to the electrostatic energy then come from the repulsions of electrons in the same orbital or from repulsions of electrons in different orbitals. The attraction arising from the third term is small. Hence, again, the molecule would behave as though the ω -regions repelled each other. Though this particular representation of χ -functions is highly idealized, it yet gives some physical insight into the forces which tend to give stability to molecules of the type XY_n , particularly in resisting deformation from the symmetrical form. Thus the first term (5.06) will in any event give the mutual repulsion of a pair of electrons in a χ -orbital, which is associated with each of the XY bonds. The interaction of bonds is due to the distribution given by the second and third terms of (5.06). The second term, being positive, will produce repulsion; the third, being negative, attraction. When the χ -functions are localized in particular bonds, it seems likely that the repulsive contribution will predominate and cause the XY bonds to be symmetrically disposed in space.

The author is indebted to Mr G. G. Hall for reading the paper in manuscript and making some valuable suggestions about notation.

REFERENCES

- Dirac, P. A. M. 1931 *Proc. Camb. Phil. Soc.* 27, 240.
 Eyring, H., Walter, J. & Kimball, G. E. 1946 *Quantum chemistry*, Appendix VII. New York: Wiley; London: Chapman & Hall.
 Lennard-Jones, J. E. 1949 *Proc. Roy. Soc. A*, 198, 1.
 Kimball, G. 1940 *J. Chem. Phys.* 8, 188.
 Mulliken, R. S. 1933 *Phys. Rev.* 43, 279.
 Pauling, L. 1931 *J. Amer. Chem. Soc.* 53, 1367.
 Slater, J. C. 1931 *Phys. Rev.* 37, 481.
 Van Vleck, J. H. & Sherman, A. 1935 *Rev. Mod. Phys.* 7, 167.

A note on polar air-mass modification

By R. FROST

(Communicated by Sir Geoffrey Taylor, F.R.S.—Received 6 October 1948—
Revised 4 February 1949)

Formulae are derived for the increase of temperature and moisture content in polar air which passes over a warm sea surface. These results are in very good agreement with observations and should be of use to forecasters.

1. INTRODUCTION

In view of the emphasis which modern methods of forecasting place on air-mass analysis the manner in which the source properties of continental polar air are modified by passing over a warm sea surface thereby absorbing heat and moisture is of fundamental importance.

The present writer in 1946 discussed the case in which air, which had been completely stirred by convection and turbulence over land to give a dry adiabatic lapse rate and a humidity mixing ratio which was constant with height, passed over a warm sea whose surface temperature was uniform, and found very good agreement between theory and observation. In practice, however, these relatively simple boundary conditions are infrequently realized. According to Sverdrup (1942), for example, when winds from the land blow over the sea the surface water is carried away from the coast and the consequent upwelling of the subsurface water which takes place brings water of greater density and lower temperature to the surface, so that in general the temperature of the sea surface increases with distance downwind from the coast, whilst observations from continental stations quoted by Petterssen (1940) show that in well-stirred continental polar air the lapse rate is generally less than the dry adiabatic, whilst the humidity mixing ratio usually decreases fairly rapidly with height.

In spite of this, many forecasters applied the theory to the practical problems of forecasting with a fair degree of success, and in an investigation carried out by staff of the Meteorological Office (of which a summary is given in an *Aviation meteorological report of eastern England*), it was found that the formulae gave good agreement with observation, the formula for the increase of temperature giving better agreement with observations than that for the increase of water vapour which tended to overestimate the increase of humidity mixing ratio.

In § 2 of the present paper the effect of a sea surface whose temperature increases uniformly with distance from the coast is discussed, and in § 3 the evaporation from an ocean surface is discussed with the assumption that the humidity mixing ratio initially decreases either linearly with height or as a power of the height. It is shown that the original formula for the increase of temperature holds to a high degree of approximation irrespective of the initial stability of the air provided that the sea temperature at the end of the trajectory is used in place of a uniform sea temperature,

and that the original formula for the increase of water vapour can be modified without difficulty to take account of an initial decrease of humidity mixing ratio with height.

2. THEORY OF TEMPERATURE INCREASE

(a) General

Let x be measured downwind from the coast and z be measured vertically upwards, then, neglecting horizontal diffusion, the equation stating that in the steady state advection and diffusion balance is

$$U \frac{\partial \theta}{\partial x} = \frac{\partial}{\partial z} \left\{ K \frac{\partial \theta}{\partial z} \right\}, \quad (2.01)$$

where U is the mean velocity of the air at a height z in the direction of x increasing, K is the coefficient of eddy diffusion and θ is the potential temperature of the air.

With U and K represented as in the earlier paper by the following conjugate power laws

$$U = U_h z^m h^{-m}, \quad (2.02)$$

$$K = m z^{1-m} z_0^{2m} U_h h^{-m}, \quad (2.03)$$

where U_h is the mean velocity of the air at a standard height h , z_0 is a length characteristic of the degree of roughness of the sea surface, and m is a non-dimensional constant which is a function of the thermal stability, equation (2.01) becomes

$$\frac{\partial \theta}{\partial x} = \alpha z^{-m} \frac{\partial}{\partial z} \left\{ z^{1-m} \frac{\partial \theta}{\partial z} \right\}, \quad (2.04)$$

where

$$\alpha = m z_0^{2m}.$$

The implications of the above power-law form of K have been discussed by the present writer (1948).

The solution of (2.04) which satisfies the boundary conditions $\theta = \theta_0$, a constant over land, and $\theta = \theta_1$, a constant over the sea, is

$$\theta = \theta_0 + (\theta_1 - \theta_0) I, \quad (2.05)$$

where

$$I = \frac{\int_{\xi}^{\infty} e^{-\xi} \xi^{-\frac{m+1}{2m+1}} d\xi}{\Gamma\left(\frac{m}{2m+1}\right)} \quad (2.06)$$

and

$$\xi = \frac{z^{2m+1}}{(2m+1)^2 \alpha x}. \quad (2.07)$$

Using a value of $m = \frac{1}{7}$ appropriate to an adiabatic lapse rate* and a value of the

* Sir Geoffrey Taylor, in a private communication, suggests that as the instability would increase with increasing distance downwind, the index m should decrease downwind, and this suggestion is supported by the observations in table 4 which show that the errors between observation and theory (with m constant) increase with increasing distance. For distances of between 100 and 500 km. which are considered in this paper, the variation of I with m is small, and no serious errors are introduced by using a constant value of the index. Even if the law of variation of m with x were known, however, the mathematical difficulties of solution of 2.04 with m varying with x would be prohibitive.

roughness parameter $z_0 = 1$ cm. appropriate to a rough sea surface as in the earlier paper (1946),

$$\xi = \frac{4.23z^{\frac{2}{3}}}{x}. \quad (2.08)$$

The values of ξ for selected values of x and z computed from equation (2.08) are given in table 1, and values of ξ for selected values of x at a height of 10 m. are shown in figure 1.

TABLE 1

$\frac{x}{z}$	1 km.	10 km.	100 km.	1000 km.
1 m.	1.6×10^{-2}	1.6×10^{-3}	1.6×10^{-4}	1.6×10^{-5}
10 m.	3.0×10^{-1}	3.0×10^{-2}	3.0×10^{-3}	3.0×10^{-4}
100 m.	5.9	5.9×10^{-1}	5.9×10^{-2}	5.9×10^{-3}

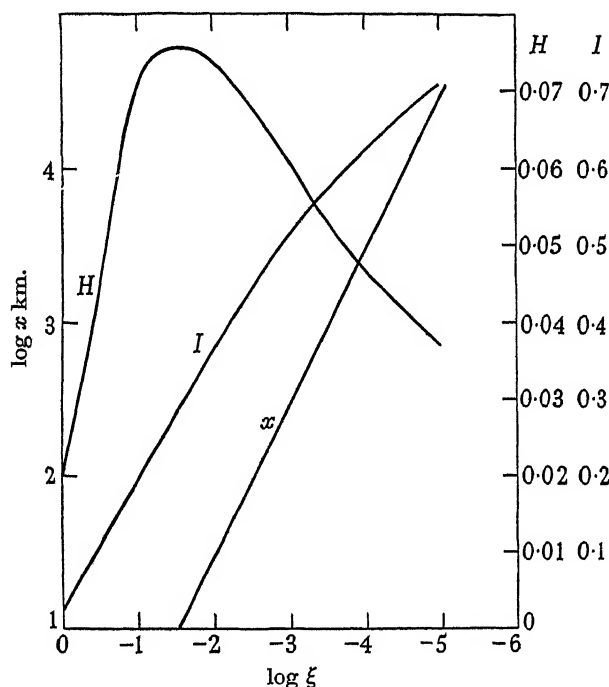


FIGURE 1. Graphical representation of tables 1 and 2.

(b) Case of $\theta = \theta_0$ over land and $\theta = \theta_1 + bx$ over the sea, where $\theta_1 \geq \theta_0$

In the present note a solution of (2.04) will first be obtained which satisfies $\theta = \theta_0$, a constant over land, and $\theta = \theta_0 + bx$, a constant over the sea, where b is a constant.

Inspection of the dimensions of the terms in (2.04) suggests that θ is a function of ξ , but a simple solution $\theta = f(\xi)$ makes θ a constant both when $\xi \rightarrow 0$ and when $\xi \rightarrow \infty$. It suggests, however, that it should be possible to find a solution

$$\theta = \theta_0 + bxf(\xi), \quad (2.09)$$

where $f(\xi) \rightarrow 1$ when $\xi \rightarrow 0$ and $f(\xi) \rightarrow 0$ when $\xi \rightarrow \infty$. (2.10)

Denoting differentiation with respect to ξ by dashed letters, $f(\xi)$ must satisfy

$$\xi f'' + \left(\xi + \frac{1+m}{1+2m} \right) f' - f = 0. \quad (2.11)$$

Writing
$$f(\xi) = \left(\xi + \frac{1+m}{1+2m} \right) \phi(\xi),$$

$\phi(\xi)$ must satisfy

$$\xi \left(\xi + \frac{1+m}{1+2m} \right) \phi'' + \left\{ 2\xi + \left(\xi + \frac{1+m}{1+2m} \right)^2 \right\} \phi' = 0, \quad (2.12)$$

and hence
$$\phi(\xi) = C \int_p^\xi \frac{e^{-\xi} \xi^{-\frac{1+m}{1+2m}}}{\left(\xi + \frac{1+m}{1+2m} \right)^2} d\xi, \quad (2.13)$$

where C and p are constants. Thus

$$f(\xi) = C \left(\xi + \frac{1+m}{1+2m} \right) \int_p^\xi \frac{e^{-\xi} \xi^{-\frac{1+m}{1+2m}}}{\left(\xi + \frac{1+m}{1+2m} \right)^2} d\xi. \quad (2.14)$$

Now from (2.10) $f(\xi) \rightarrow 0$ when $\xi \rightarrow \infty$, and hence $p = \infty$. Therefore

$$f(\xi) = -C \left(\xi + \frac{1+m}{1+2m} \right) \int_\xi^\infty \frac{e^{-\xi} \xi^{-\frac{1+m}{1+2m}}}{\left(\xi + \frac{1+m}{1+2m} \right)^2} d\xi. \quad (2.15)$$

Integrating by parts twice

$$f(\xi) = C \left(\xi + \frac{1+m}{1+2m} \right) \left[\left(\frac{1+2m}{1+m} \right) \frac{e^{-\xi} \xi^{\frac{m}{1+2m}}}{\left(\xi + \frac{1+m}{1+2m} \right)} - \frac{1+2m}{1+m} \int_\xi^\infty e^{-\xi} \xi^{-\frac{1+m}{1+2m}} d\xi \right]. \quad (2.16)$$

Since from (2.10) $f(\xi) \rightarrow 1$ when $\xi \rightarrow 0$

$$-C \Gamma \left(\frac{m}{1+2m} \right) = 1, \quad (2.17)$$

and therefore
$$f(\xi) = \left(1 + \frac{1+2m}{1+m} \xi \right) I - \left(\frac{1+2m}{1+m} \right) \frac{e^{-\xi} \xi^{\frac{m}{1+2m}}}{\Gamma \left(\frac{m}{1+2m} \right)}. \quad (2.18)$$

Hence
$$\theta = \theta_0 + bx \left[\left(1 + \frac{1+2m}{1+m} \xi \right) I - \left(\frac{1+2m}{1+m} \right) \frac{e^{-\xi} \xi^{\frac{m}{1+2m}}}{\Gamma \left(\frac{m}{1+2m} \right)} \right] \quad (2.19)$$

is the solution which satisfies the required conditions.

Since both (2.05) and (2.19) are solutions of (2.04), it follows that

$$\theta = \theta_0 + (\theta_1 - \theta_0) I + bx \left[\left(1 + \frac{1+2m}{1+m} \xi \right) I - \frac{(1+2m) e^{-\xi} \xi^{\frac{m}{1+2m}}}{(1+m) \Gamma \left(\frac{m}{1+2m} \right)} \right] \quad (2.20)$$

is also a solution of (2.04), and it can readily be seen that (2.20) is the solution of the differential equation which fits the boundary conditions $\theta = \theta_0$ over land and $\theta = \theta_1 + bx$ over the sea where $\theta_1 \geq \theta_0$.

Writing $\theta_1 + bx = \theta_2$, equation (2.20) may be written

$$\theta = \theta_0 + (\theta_2 - \theta_0)I + \left(\frac{1+2m}{1+m}\right)(\theta_2 - \theta_1) \left\{ \xi I - \frac{e^{-\xi} \xi^{\frac{m}{1+2m}}}{\Gamma\left(\frac{m}{1+2m}\right)} \right\} \quad (2.21)$$

or
$$\theta = \theta_0 + (\theta_2 - \theta_0)I - (\theta_2 - \theta_1)H, \quad (2.22)$$

where
$$H = \left(\frac{1+2m}{1+m}\right) \left\{ \frac{e^{-\xi} \xi^{\frac{m}{1+2m}}}{\Gamma\left(\frac{m}{1+2m}\right)} - \xi I \right\}. \quad (2.23)$$

Values of H for selected values of ξ , together with the values of I which have been extracted from table 1 of the earlier paper, are given in table 2 and are shown graphically in figure 1.

TABLE 2

ξ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1
H	0.037	0.047	0.061	0.074	0.071	0.019
I	0.706	0.621	0.510	0.368	0.190	0.027

Since $\theta_1 \geq \theta_0$, it follows that for very small values of ξ , the temperature is given to a good degree of approximation by

$$\theta = \theta_0 + (\theta_2 - \theta_0)I, \quad (2.24)$$

which is similar to (2.05) except that θ_2 , the temperature of the sea at a distance x downwind, now replaces θ_1 , a uniform sea-surface temperature.

At any distance x downwind it follows from equation (2.22) that the potential temperature θ can be expressed by an equation of the type

$$\theta = \theta_0 + (\bar{\theta} - \theta_0)I, \quad (2.25)$$

where $\bar{\theta}$ is the weighted mean of the initial and final sea temperature and is given by

$$\bar{\theta} = \frac{\theta_2(I - H) + \theta_1 H}{I}. \quad (2.26)$$

Table 3, which is derived from figure 1, gives values of I and H at a height of 10 m. at various distances downwind from the coast.

TABLE 3

distance in km.	10	50	100	200	300
I	0.28	0.40	0.44	0.48	0.51
H	0.076	0.071	0.070	0.064	0.061
distance in km.	400	500	600	800	1000
I	0.525	0.54	0.55	0.56	0.57
H	0.059	0.057	0.056	0.055	0.054

From table 3 it can be seen that for calculating the potential temperature at a height of say 10 m., the final and initial sea temperatures have to be weighted in the ratio $2\frac{1}{2}:1$ at a distance of about 10 km. from the coast, in the ratio of 4:1 at a distance of about 40 km. from the coast, in the ratio 7:1 at a distance of about 200 km. from the coast, and in the ratio of 10:1 at a distance of about 1100 km. For most practical purposes therefore it should be sufficient to know the temperature of the sea at the end of the trajectory where fortunately it can most readily be measured.

(c) *Comparison with observation*

Klein, Espy & Palladino (1945), in an investigation into methods of forecasting temperatures on board weather ships during outbreaks of polar air over the north Atlantic, when climatological data showed that the increase of sea temperature with distance from the coast was approximately linear, found that ΔT (the temperature difference between the air temperature as measured at a weather ship and the temperature of the sea surface in the vicinity of the ship) was more highly correlated with the sea-surface temperature in the vicinity of the ship than with the average sea-surface temperature between the land and the ship. Burke (1945), who developed a method of forecasting temperatures on board ship during such outbreaks of polar air, for which a knowledge of the sea-surface temperature was required, found for trajectories of less than 680 km. over the sea that close agreement could be obtained by using a mean sea-surface temperature in which the final sea-surface temperature was weighted three times as heavily as the initial sea-surface temperature, whilst for trajectories of 680 km. or over he found that the closest agreement was obtained by using the final sea-surface temperature only.

In table 4 a comparison between observations given by Burke (1945) and values calculated from equation (2.24) with the aid of table 3 is given. The agreement is reasonably good.

TABLE 4

length of over-water trajectory in km.	initial surface air temperature in ° K	average lapse rate over land between 0 and 1500 m. as % of dry adiabatic lapse rate	mean or final sea- surface temperature	observed air tem- perature	calculated air tem- perature	error	error by Burke's method
240	269	85	276	273	272.5	- 0.5	0
870	262	73	276	269	270.0	+ 1.0	+ 2.0
1020	268	80	279	272	274.5	+ 2.5	+ 4.0

In the majority of the observations given by Burke however, the average lapse rate in the cold air over the land differed markedly from the dry adiabatic lapse rate and the effect of this upon the final temperature is difficult to assess quantitatively. The effects of any initial stratification of the air are (i) to damp down the eddies and thereby to reduce the flux of heat from the sea to the air and (ii) to confine this reduced heat within a smaller thickness of the atmosphere, and these effects are in opposition. This can readily be seen from a diagram. Thus in figure 2 curves a' and b' represent the variation of temperature with pressure at a given distance downwind over the sea corresponding to the initial states over the land, a = a dry adiabatic

lapse rate and b = a stable lapse rate. If the flux of heat from the sea to the air were the same in the two cases the areas enclosed by curves aa' and the $\log p$ axis and curves bb' and the $\log p$ axis would be equal. The temperature of the air near the surface at a given distance downwind over the sea would therefore be greater when the air is initially stable than when it is initially in neutral equilibrium. As, however, the effect of any initial stratification is also to reduce the flux of heat, the area between bb' and the $\log p$ axis is less than the area between aa' and the $\log p$ axis, and hence the difference between the final air temperatures is correspondingly reduced. It might therefore be expected that over a limited distance of travel downwind over the sea, the distance increasing with proximity to the sea surface, formulae (2.22) and (2.24) would give a good approximation to the observed temperatures irrespective of the initial stability of the air.

TABLE 5

length of over-water trajectory x	initial surface-air tempera- ture θ_0	initial lapse rate as % of dry adiabatic lapse rate	initial sea- surface tempera- ture θ_1	final sea- surface tempera- ture θ_2	Burke's mean sea- surface tempera- ture $\frac{3\theta_1 + \theta_2}{4}$	observed air tempera- ture	calculated air tempera- ture	error from equation (2.24)	error from equation (2.22)
240	269.0	85	—	—	276	273	272.5	-0.5	—
320	274.0	53	—	—	277	276	274.5	-1.5	—
330	260.0	47	—	—	274	268	267	-1.0	—
355	274.5	47	281.5	283	—	277.5	279	—	+1.5
370	265.0	0	281.5	289.5	—	279.0	277.5	—	-1.5
400	274.0	40	—	—	279	278	276.5	-1.5	—
420	267.0	36	277	284	—	276	275.5	—	-0.5
425	273.0	13	—	—	277	274	275	1.0	—
430	268.5	31	277	284	—	276	276.5	—	+0.5
440	271.0	29	284	287	—	278.5	279.5	—	+1.0
490	267.5	27	282.5	292.5	—	279.0	280.5	—	+1.5
500	266.5	13	282	285	—	277.0	276.5	—	-0.5
500	264.0	25	282	291	—	278	278	—	0.0
505	268.0	53	—	—	274	272	271	-1.0	—
520	263.5	28	277.5	284	—	275.5	274	—	-1.5
550	263.5	30	277.5	284.5	—	275	274.5	—	-0.5

Table 5 shows that for a height of 10 m. above the surface of the sea Burke's observations support this expectation for distances of travel up to 550 km. from the coast. It can be seen from this table that in the cases for which both the initial and final sea-surface temperatures are available and it is possible to use the more exact equation (2.22), the mean error is 0° K and the mean absolute error is 0.9° K. In the other cases for which, in the absence of a final sea-surface temperature, Burke's mean sea temperature has been used, the mean error is -0.75° K, which suggests that Burke has given too much weight to the initial sea-surface temperature in obtaining the mean.

For distances greater than 550 km. comparison of Burke's observations with temperatures calculated from equation (2.22) or (2.24) shows that the calculated values are too low. The errors are roughly proportional to the difference between the actual lapse rate and the dry adiabatic lapse rate and increase with increasing distance over 500 km.

For forecasting temperature at a height of 1.2 m. (4 ft.) it is probable that formulae (2.22) and (2.24) would give a good approximation to the observed temperature for trajectories of more than 1000 km.

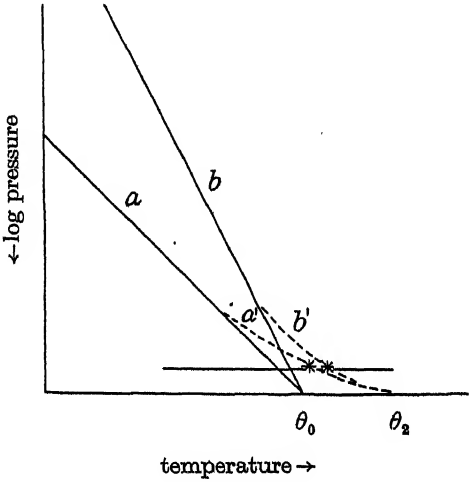


FIGURE 2. Temperature increase due to initial stability illustrated diagrammatically.

3. THEORY OF MOISTURE INCREASE

(a) General

The partial differential equation (2.04) may be used for the discussion of the flux of water vapour providing θ , the potential temperature in the equation, is replaced by μ , the humidity mixing ratio, i.e.

$$\frac{\partial \mu}{\partial x} = \alpha z^{-m} \frac{\partial}{\partial z} \left(z^{1-m} \frac{\partial \mu}{\partial z} \right), \tag{3.01}$$

where $\alpha = m z_0^{2m}$.

The solution of the equation which satisfies the boundary conditions $\mu = \mu_0$, a constant over land, and $\mu = \mu_s$, a constant over the sea, is as in (2.05)

$$\mu - \mu_0 = (\mu_s - \mu_0) I, \tag{3.02}$$

where I is given by (2.06). Using as before a value of $m = \frac{1}{7}$ for unstable lapse rates and a value of the roughness parameter $z_0 = 1$ cm. appropriate to a rough sea surface,

TABLE 6

x (km.)	...	100	200	300	400
ξ		2×10^{-4}	10^{-4}	6.7×10^{-5}	5×10^{-5}
I		0.598	0.620	0.637	0.650
$\xi^{-\frac{1}{2}}$		2.93×10^2	4.64×10^2	6.07×10^2	7.35×10^2
x (km.)	...	500	700	900	1000
ξ		4×10^{-5}	2.8×10^{-5}	2.2×10^{-5}	2×10^{-5}
I		0.660	0.671	0.679	0.682
$\xi^{-\frac{1}{2}}$		8.57×10^2	1.08×10^3	1.27×10^3	1.36×10^3

the values of I and ξ at a height of 4 ft. for selected values of x computed from equations (2.06) and (2.07) are given in table 6. The last line gives values of $\xi^{-\frac{1}{2}}$ which are required at a later stage.

(b) Case of $\mu = \mu_0 - \lambda z$ over the land and $\mu = \mu_s$ a constant over the sea

A solution of equation (3.01) will now be derived which satisfies the boundary conditions $\mu = \mu_s$ over the sea and $\mu = \mu_0 - \lambda z$ over the land.

Inspection of the dimensions of the terms in (3.01) suggests that μ is a function of ξ , but a simple solution $\mu = f(\xi)$ makes μ a constant both when $\xi \rightarrow 0$ and when $\xi \rightarrow \infty$. It suggests, however, that it should be possible to find a solution

$$\mu - \mu_0 = (\mu_s - \mu_0) I - \lambda z f(\xi). \quad (3.03)$$

Denoting differentiation with respect to ξ by dashed letters, $f(\xi)$ must satisfy

$$\xi^2 f''(\xi) + \left(\xi^2 + \frac{3+m}{1+2m} \xi \right) f'(\xi) + \frac{1-m}{(1+2m)^2} f(\xi) = 0. \quad (3.04)$$

Writing

$$f(\xi) = \phi(\xi) e^{-\frac{1}{2}\xi} \xi^{-\frac{3+m}{2(1+2m)}},$$

then $\phi(\xi)$ must satisfy

$$\phi''(\xi) + \phi(\xi) \left[-\frac{1}{4} - \frac{3+m}{2(1+2m)\xi} + \frac{1 - \left(\frac{m}{1+2m} \right)^2}{4\xi^2} \right] = 0. \quad (3.05)$$

This is, however, the confluent hypergeometric (Whittaker & Watson 1927, chapter 16) equation of which two independent solutions are

$$M_{k,n}(\xi) = \xi^{\frac{1}{2}+n} e^{-\frac{1}{2}\xi} \left\{ 1 + \frac{\frac{1}{2}+n-k}{1!(1+2n)} \xi + \frac{(\frac{1}{2}+n-k)(\frac{3}{2}+n-k)}{2!(1+2n)(2+2n)} \xi^2 + \dots \right\}, \quad (3.06)$$

and $M_{k,-n}(\xi) = \xi^{\frac{1}{2}-n} e^{-\frac{1}{2}\xi} \left\{ 1 + \frac{\frac{1}{2}-n-k}{1!(1+2n)} \xi + \frac{(\frac{1}{2}-n-k)(\frac{3}{2}-n-k)}{2!(1-2n)(2-2n)} \xi^2 + \dots \right\}, \quad (3.07)$

where $k = -\frac{3+m}{2(1+2m)}$ and $n = \frac{m}{2(1+2m)}, \quad (3.08)$

and hence the general solution is

$$\phi(\xi) = A M_{k,n}(\xi) + B M_{k,-n}(\xi), \quad (3.09)$$

where A and B are arbitrary constants which have to be determined by the boundary conditions.

Now since when $z \rightarrow 0$, $\xi \rightarrow 0$ and hence $\mu \rightarrow \mu_s$, it follows that

$$\lim_{z \rightarrow 0} \lambda z \xi^{-\frac{3+m}{2(1+2m)}} e^{-\frac{1}{2}\xi} \{ A M_{k,n}(\xi) + B M_{k,-n}(\xi) \} = 0, \quad (3.10)$$

or $\lim_{z \rightarrow 0} \lambda z \xi^{-\frac{3+m}{2(1+2m)}} \left\{ A \xi^{\frac{1}{2} + \frac{m}{2(1+2m)}} + B \xi^{\frac{1}{2} - \frac{m}{2(1+2m)}} \right\} = 0, \quad (3.11)$

or $\lim_{z \rightarrow 0} \lambda \left[A z^m \{ (2m+1)^2 a x \}^{\frac{1-m}{1+2m}} + B \{ (2m+1)^2 a x \}^{\frac{1}{1+2m}} \right] = 0, \quad (3.12)$

whence $B = 0$.

Now when $x \rightarrow 0$, $\xi \rightarrow \infty$ and $\mu \rightarrow \mu_0 - \lambda z$, and hence

$$\lim_{\xi \rightarrow \infty} \xi^{-\frac{3+m}{2(1+2m)}} e^{-\frac{1}{2}\xi} A M_{k,n}(\xi) = 1, \quad (3.13)$$

$$\lim_{\xi \rightarrow \infty} A \xi^{-\frac{3+m}{2(1+2m)}} e^{-\frac{1}{2}\xi} \left[\frac{\Gamma(1+2n)}{\Gamma(\frac{1}{2}+n-k)} e^{\pi i k} W_{-k,n}(-\xi) + \frac{\Gamma(1+2n)}{\Gamma(\frac{1}{2}+n+k)} e^{(\frac{1}{2}+n+k)\pi i} W_{k,n}(\xi) \right] = 1, \quad (3.14)$$

where $W_{k,n}(\xi)$ is the Whittaker function, and making use of the asymptotic expansions for $W_{k,n}(\xi)$ and $W_{-k,n}(-\xi)$,

$$\lim_{\xi \rightarrow \infty} A \xi^{-\frac{3+m}{2(1+2m)}} e^{-\frac{1}{2}\xi} \left[\frac{\Gamma(1+2n)}{\Gamma(\frac{1}{2}+n-k)} e^{\pi i k} e^{\frac{1}{2}\xi} (-\xi)^{-k} + \frac{\Gamma(1+2n)}{\Gamma(\frac{1}{2}+n+k)} e^{(\frac{1}{2}+n+k)\pi i} e^{-\frac{1}{2}\xi} \xi^k \right] = 1, \quad (3.15)$$

$$\text{whence} \quad A \frac{\Gamma(1+2n)}{\Gamma(\frac{1}{2}+n-k)} = 1 \quad (3.16)$$

$$\text{or} \quad A = \frac{\Gamma(\frac{1}{2}+n-k)}{\Gamma(1+2n)} = \frac{\Gamma\left(1 + \frac{1}{1+2m}\right)}{\Gamma\left(1 + \frac{m}{1+2m}\right)}, \quad (3.17)$$

and hence the solution of equation (3.01) which satisfies the required boundary conditions is

$$\mu - \mu_0 = (\mu_s - \mu_0) I - \lambda z e^{-\frac{1}{2}\xi} \xi^{-\frac{3+m}{2(1+2m)}} \frac{\Gamma\left(1 + \frac{1}{1+2m}\right)}{\Gamma\left(1 + \frac{m}{1+2m}\right)} M_{k,n}(\xi). \quad (3.18)$$

For distances downwind from the coast of over 100 km. and for small heights above the surface of the sea ξ is very small, and hence from (3.06) μ may be written to a high degree of approximation by

$$\mu - \mu_0 = (\mu_s - \mu_0) I - \lambda z \frac{\Gamma\left(1 + \frac{1}{1+2m}\right)}{\Gamma\left(1 + \frac{m}{1+2m}\right)} \xi^{-\frac{1-m}{1+2m}}. \quad (3.19)$$

(c) *Power-law decrease of humidity mixing ratio with height*

If the initial humidity mixing ratio instead of decreasing linearly with height, decreased as a power of the height, i.e. $\lim_{x \rightarrow 0} \mu = \mu_0 - cz^p$, where $m \leq p \leq 1$, a similar analysis to the above gives

$$\mu - \mu_0 = (\mu_s - \mu_0) I - cz^p e^{-\frac{1}{2}\xi} \xi^{-\frac{1+2p+m}{2(1+2m)}} \frac{\Gamma\left(1 + \frac{p}{1+2m}\right)}{\Gamma\left(1 + \frac{m}{1+2m}\right)} M_{k,n}(\xi), \quad (3.20)$$

$$\text{where} \quad k = -\frac{1+2p+m}{2(1+2m)} \quad \text{and} \quad n = \frac{m}{2(1+2m)}, \quad (3.21)$$

and the approximate solution in this case is

$$\mu - \mu_0 = (\mu_s - \mu_0) I - cz^p \frac{\Gamma\left(1 + \frac{p}{1+2m}\right)}{\Gamma\left(1 + \frac{m}{1+2m}\right)} \xi^{-\frac{(p-m)}{1+2m}}. \quad (3.22)$$

Application

If in (3.19) m is given a value of $\frac{1}{2}$, then

$$\mu - \mu_0 = (\mu_s - \mu_0) I - \lambda z \times 0.98 \xi^{-\frac{1}{2}}, \quad (3.23)$$

and hence the effect of an initial humidity mixing ratio which decreases linearly with height is to decrease the humidity mixing ratio as given by the simple rule (3.02) by approximately

$$\lambda z \xi^{-\frac{1}{2}}. \quad (3.24)$$

Hann (1929) gave the following formula for the average variation of vapour pressure with height:

$$e = e_0 10^{-\frac{z}{630003}}, \quad (3.25)$$

where e_0 is the vapour pressure at the ground and z is measured in cm.

This formula for heights up to about 1000 m. gives results not very different from the formula adopted by Kaminski (1900) in the *Climatological atlas of the Russian Empire* for the reduction of values of vapour pressure to sea-level,

$$e_0 = e(1 + 4 \times 10^{-6}z), \quad (3.26)$$

and since the vapour pressure near the surface is very nearly proportional to the humidity mixing ratio, either formula gives to a very good degree of approximation

$$\mu = \mu_0(1 - 4 \times 10^{-6}z), \quad (3.27)$$

which is the initial condition assumed in the discussion of this problem. Thus

$$\lambda z = \mu_0 \times 4 \times 10^{-6}z. \quad (3.28)$$

At a height of 4 ft. above the surface therefore the humidity mixing ratio as given by the simple formula (3.02) is in excess of the actual humidity mixing ratio by approximately

$$5 \times 10^{-4} \times \mu_0 \xi^{-\frac{1}{2}}. \quad (3.29)$$

In a case discussed in the *Aviation meteorological report for eastern England*, during a spell of 'easterlies' in January 1941, air left the continent with a temperature and dew-point of 21 and 17° F respectively, and passed over the North Sea whose surface temperature was 42° F. The observed values of the temperature of the air arriving at the east coast varied from 34 to 35° F and the dew-point from 30 to 31° F. Now from equation (2.05) and table 6 the calculated temperature after a trajectory of 400 km. is

$$T = 21 + 0.65(42 - 21) = 34.6^\circ \text{F},$$

in good agreement with the observed values. Similarly from equation (3.02) and table 6 the calculated humidity mixing ratio corresponding to the above temperature is

$$\mu = 1.9 + 0.65(5.65 - 1.9) = 4.33,$$

which corresponds to a dew-point of 35.0°F which is clearly far too high. If, however, allowance is made for the initial decrease of humidity mixing ratio with height, then from equations (3.02) and (3.29), the humidity mixing ratio is given by

$$\begin{aligned}\mu &= 1.9 + 0.65(5.65 - 1.9) - 1.9 \times 5 \times 10^{-4} \times 7.35 \times 10^2 \\ &= 4.33 - 0.7 \\ &= 3.63,\end{aligned}$$

which corresponds to a dew-point of 30.7°F , in very good agreement with the observed values.

It can be seen from equation (3.29) that for a distance of travel of 375 km. the correction for the initial decrease of humidity mixing ratio with height is $0.35\mu_0$, and hence for this distance the final humidity mixing ratio is

$$\begin{aligned}\mu &= 0.65\mu_s + 0.35\mu_0 - 0.35\mu_0 \\ &= 0.65\mu_s,\end{aligned}$$

or in other words the dew-point of the air at 4 ft. at this distance is independent of the initial dew-point of the air and depends only upon the sea temperature.

It is considered that formula (3.02), as modified by (3.29), should provide a valuable aid to forecasters.

I am indebted to the Director of the Meteorological Office for permission to publish this paper.

REFERENCES

- Burke, C. J. 1945 *J. Met., Milton (Mass.)*, 2, no. 2.
 Frost, R. 1946 *Proc. Roy. Soc. A*, 186, 20.
 Frost, R. 1948 *Quart. J. R. Met. Soc.* 74, 316.
 Hann, J. 1929 *Lehrbuch der meteorologie*, 4 Auflage. Leipzig: Chr. Herm. Touchnitz.
 Kaminski, A. 1900 *Atlas Climatologique de l'Empire de Russie*. St Petersburg: l'observatoire Physique Central Nicolas.
 Klein, W. H., Espy, R. B. & Palladino, A. J. 1945 *U.S.A.A.F. Rep.* no. 971, Washington.
 Petterssen, S. 1940 *Weather analysis and forecasting*, p. 179. New York: McGraw-Hill Book Co.
 Sverdrup, H. U. 1942 *Oceanography for meteorologists*. New York: Prentice Hall Inc.
 Whittaker, E. T. & Watson, G. N. 1927 *A course of modern analysis*. Cambridge University Press.

Unified field theory in a curvature-free five-dimensional manifold

By J. G. BENNETT, R. L. BROWN AND M. W. THRING

Institute for the Comparative Study of History, Philosophy and the Sciences, Kingston, Surrey.

(Communicated by N. F. Mott, F.R.S.—Received 3 November 1948)

Instead of identifying fields with the curvature of a metric, the present theory shows that they may be identified with the manner in which the four-way measuring system of the physical observer O is embedded in a flat five-dimensional manifold provided that due account is taken of the imperceptibility of the fifth dimension. In this system fields are introduced by treating the direction cosines, ${}^k l_\mu$, of the four directions of measurement and of the imperceptible direction as variable functions of position in the manifold. The track of an unconstrained body P is taken as a straight line (cosmodesic) in the manifold, but the 'projection' of it which O observes in his four-co-ordinate system is in general curved. Thus the equation describing the element ds of P 's cosmodesic in O 's four-co-ordinate system (Δx^μ) is

$$ds^2 \cos^2 \lambda - 2ds \sin \lambda \left\{ \left(\sum_{\nu=1}^4 {}^\nu l_\mu {}^\nu l_\mu \right) \Delta x^\mu \right\} = \left\{ {}^5 l_\mu {}^5 l_\nu - 2 {}^5 l_\nu \sum_{k=1}^5 {}^k l_\mu {}^k l_\nu + \sum_{k=1}^5 {}^k l_\mu {}^k l_\nu \right\} \Delta x^\mu \Delta x^\nu.$$

When O applies the variational condition to ds which expresses the fact that the cosmodesic is straight, he concludes that it has a space-time curvature with two distinct components, one dependent upon λ which is the angle between the cosmodesic and an universal direction ${}^5 Q$ and upon ${}^\nu l_\mu$, the other acting equally on all P bodies whatever the value of λ and depending only on ${}^5 l_\mu$. These 'accelerations' are shown to correspond to electromagnetic and gravitational fields respectively, and the inverse square law of force is shown to hold for spherically symmetrical fields of both types as a consequence of the condition of coherence of the measuring system.

When the cause of the positional variation of the ${}^k l_\mu$ is a heavy body, having a constrained rotation, it is shown to give rise to the magnetic field that a body of charge equal to its gravitational mass would have, without the corresponding electrostatic field.

The ${}^k l_\mu$'s are restricted by the requirement that the angles between the absolute fifth direction, the direction imperceptible for O , and the direction orthogonal to O 's four measuring directions, are all null.

A list of symbols used in this paper is given in an appendix on p. 60.

1. GEOMETRICAL REPRESENTATION OF FREEDOM AND CONSTRAINT

1.1. *Types of field theory*

Field theories are framed to give a mathematical account of the observed curvilinear paths of unconstrained moving bodies. There are various ways in which such theories can be constructed. A primary distinction can be drawn according to whether the paths are regarded as 'really' curved, or 'really' straight. If they are 'really' curved, they may be described in terms of central forces 'acting at a distance'. This is the Newton-Maxwell type of theory, and it fails to give a complete account of the observed facts. Alternatively, the curvature can be ascribed to the metric framework to which the observations are referred. This leads to so-called 'geometrical theories'. These can be constructed to give an account, accurate within the present

limits of verifiability, of unconstrained gravitational motions. They fail, however, to offer any simple and readily visualized representation of such motions, and still less do they make it easy to represent the electromagnetic field or the connexion between these two disclosed by the gravitomagnetic effect.

A third line is to take the paths of all unconstrained bodies as absolutely straight—thus adopting a simple and natural extension of Newton's first law. In this case it is necessary to postulate an absolute reference manifold, itself free from curvature, without which absolute straightness has no meaning. Both the path and the reference manifold being regarded as free from curvature, the observed curvature of the motion must, in this type of theory, be ascribed solely to some property of the observer and his measuring system and the manner in which this system is embedded in the reference manifold. In the present paper a theory of the 'absolute straight path' type is developed in which the reference manifold is five-dimensional and curvature-free. This appears to conflict with the demonstration that the geometrical interpretation of the gravitational field requires a non-Euclidean metric (Schwarzschild equation) that cannot be embedded in a Euclidean manifold of less than ten dimensions. (Eddington 1924). It will, however, be shown that this difficulty does not arise if the path of the unconstrained body and the measuring system of the observer are independently related to the reference manifold.

1.2. *Statement of basic postulates*

In order to give a mathematical account of the results of identifying the curvature of observed paths with properties of the observer and his measuring system it is necessary to express these properties in a geometrical form. They are: as regards the observer, the fact that his measurements are, at any given point, confined to only four (three of space and one of time) out of the possible five independent directions in the cosmic manifold; and, as regards the measuring system, the fact that any situation in the five-dimensional manifold must, by a generalized projection, be referred back to the measuring system. The rigidity of O 's measuring system enables the directions of measurement at two spatially separate points to be related to one another. The manner in which this is effected, combined with the restriction upon the measurements of the observer, gives rise to the equations for the field. In this way, a theory is constructed which is entirely free from action at a distance, a result which has not been obtained even in geometrical as opposed to central force theories.

These considerations can be formulated in terms of postulates defining four constituent elements in a generalized dynamical system.

(a) A five-dimensional flat cosmic manifold and an 'absolute observer' Q , able to measure 'true' intervals RS between any two points in the manifold. If then Q sets up a rectangular, orthogonal co-ordinate system (${}^1Q, {}^2Q, {}^3Q, {}^4Q, {}^5Q$) for the whole manifold, with origin R , the interval RS will satisfy the relation

$$RS^2 = -({}^1Q)^2 - ({}^2Q)^2 - ({}^3Q)^2 + ({}^4Q)^2 + ({}^5Q)^2. \quad (1.1)$$

In this equation the kQ , the co-ordinates in the Q system of the point S , are all real numbers and the difference between the space-like (${}^1Q, {}^2Q, {}^3Q$) and time-like

(${}^4Q, {}^5Q$) directions is indicated by the Minkowskian device of positive and negative additions of the squares of these numbers. The fifth (time-like) co-ordinate can conveniently be called 'anti-time' or 'eternity'. The term anti-time will be used in the present paper. The property of irreversibility associated with temporal processes does not arise in field theory, nevertheless it may be worth noting that the anti-time postulate suggested itself to us from the thermodynamical consideration that conservation and irreversibility hold only for temporal processes. Symmetry suggests that there should also be a direction along which these effects are reversed and we find in fact that for anti-time displacements entropy is conserved but the total energy content of a closed system may have different values.

(b) A physical observer O who is 'anti-time blind'; i.e. who can only make measurements of space-like intervals (rigid rulers) and time-like intervals (clocks) but can neither observe nor make measurements in a certain direction which is near to the universal anti-time direction 5Q .^{*} O depends for his observations upon his own rigid measuring system, and it is assumed that, referred to the cosmic manifold, this need be neither rectilinear nor orthogonal at any given point. We shall describe the kind of time constancy of shape and size, possessed by O 's constructional materials, by the term *relative rigidity* to distinguish it from *absolute rigidity* which would apply to configurations self-congruent for all possible linear transformations in the cosmic manifold. We can readily see that the relative rigidity of O 's experience does not exist for Q . Let R be taken as origin of O 's co-ordinate system. The measurements given by O 's rulers (X^1, X^2, X^3) and clocks (X^4) starting from the point R are thus taken along curved lines through the point; these are assumed to lie close to but not necessarily to coincide with the rectilinear orthogonal axes of Q at the point R . Thus what appears to be rigid for O changes shape for Q . This implies that the distance between two of O 's time lines, which for him is constant, is not in general constant for Q . It is by the conditions which have to be introduced to ensure the stationary time character and isotropic space character of O 's measuring system that the conception of 'restricted rigidity' is given a geometrical interpretation.

These conditions can be formulated in terms of the conservation principle that for any possible observer a stationary system does no work. An equivalent, if less obvious, statement is that the density of a body on which no work is done remains constant for any observer. From this it follows that the volume of a 'relatively rigid' body (e.g. O 's ruler) remains stationary for Q , although by Q 's measurements, the distance between any two points changes with O 's time. In other words, the ruler which is ideally rigid for O behaves as if it were ideally plastic for Q . As seen by Q the dimensions of an element of a rigid body change in such a way that the alteration in the direction of the field-producing body compensates those in the orthogonal directions so that the volume remains constant. For O , of course, the dimensions do not change.

(c) An unconstrained body P , the observation of which is the only means available to the observer for detecting the presence of a field. P is without space extension,

^{*} It will appear later that besides being small the angle between these two directions must be a null angle, i.e. an angle whose cosine is unity.

and its temporal existence can be represented by a straight line in the Q -co-ordinate system. To indicate the 'absolute' character of the straightness of P 's path we shall use the term 'cosmodesic'. To demonstrate the fields it is sufficient to show that the apparent track for O of a P whose cosmodesic passes through the origin R has the appropriate curvature.

(d) A field-producing system M whose space-distributed existence is the non-vanishing of the components of null angles between 5Q , ϖ and L_5 at all points in the cosmic manifold. M is not necessarily rigidly connected to O , and the null angle may be composed of several simple 'fields' centred at a set of points occupying a definite volume of space, i.e. M may be of finite size or a point.

Field theory then becomes the science of the relations between simple unconstrained point bodies moving in cosmodesics and the space-extended rigid systems used by physical observers for making measurements.

1.3. Notation and conventions

It is convenient to replace the purely real (in the mathematical sense) sets of numbers kQ and X^μ by the mixed sets* defined by

$${}^\psi q \equiv i{}^\psi Q, \quad {}^4q \equiv {}^4Q, \quad {}^5q \equiv {}^5Q; \quad x^\psi \equiv iX^\psi, \quad x^4 \equiv X^4 \quad (\psi = 1, 2, 3). \quad (1.1a)$$

1.4. Universal anti-time

We assume that the general mass-system of the universe determines an unique direction of anti-time 5Q in the cosmic manifold. All 5Q 's at all times and in all places are therefore parallel. From this it follows that the cosmodesic of an unconstrained body P always makes a fixed angle with 5Q , so that we can take an angle λ such that $\frac{1}{2}\pi - \lambda$ is the angle made by the cosmodesic of P with 5Q . 4Q is the direction of time for Q , and the choice of the 4Q direction from among all possible time-like directions in the four world through R orthogonal to 5Q is equivalent to fixing the velocity of Q relative to the mass-system of the universe. Owing to the isotropy of space 1Q , 2Q and 3Q can be arbitrarily selected as any three imaginary axes orthogonal to 4Q and 5Q .

1.5. The rigid measuring system of O

We now define the measuring system of O by means of five directions at any point (e.g. the origin R), having vector components nearly equal to those of the kQ . These five directions will be designated by the unit vectors L_1, L_2, L_3, L_4, L_5 , where L_4 is the direction at R along which O measures time; L_5 is the direction at R along which displacement involves no change in the clock and ruler readings of O . There can be only one such direction because the manifold is five-dimensional and O can make four independent measurements. Ruler and clock measurements made

* The summation convention will be adopted as follows:

indices	j, k, n, s, v, w, p	take values	1, 2, 3, 4, 5,
	μ, ν, σ	take values	1, 2, 3, 4,
	ψ, η, ξ	take values	1, 2, 3.

Indices will be written as affixes when they refer to Q and as suffices when they refer to O . Thus ${}^k l_j$ signifies the cosine of the angle between kQ and the L_j at the same point where L_j is a vector associated with O .

along L_μ will be expressed by the symbol Δx^μ , where $\mu = 1, 2, 3$ and 4 , but not 5 . We also introduce at the origin R a direction ϖ having unit vector components in the Q -system ${}_1\varpi, {}_2\varpi, {}_3\varpi, {}_4\varpi, {}_5\varpi$; ϖ like L_5 , makes a small angle with 5Q , but it is defined as being orthogonal to the four vectors L_1, L_2, L_3, L_4 at R . We have thus at each point of the cosmic manifold accessible to O three anti-time-like directions (1) 5Q which is universal, (2) L_5 and (3) ϖ , the latter two being local and serving to fix the rigidity of O 's measuring system.

1.6. Null intervals and null angles

It is an obvious property of a complex geometry that it is possible to have null intervals

$$RS = \sqrt{(-{}^1Q^2 - {}^2Q^2 - {}^3Q^2 + {}^4Q^2 + {}^5Q^2)} = 0, \quad (1.2)$$

where the Q are all real and finite.

Similarly, we call the relation between a pair of unit vectors a 'null angle' when they have the same or different components and the cosine of the angle between them, defined in the usual way as their scalar product, is unity. The angle will be called a 'zero angle' when all the components are identical.

1.7. Gravitational field

The gravitational field is now defined as the situation which satisfies the condition that at all points in the manifold L_5 coincides with 5Q and ϖ makes a null angle with 5Q with non-zero components that are themselves functions of kQ . In other words

$$\mu l_5 = 0, \quad {}^5l_5 = 1, \quad (1.3)$$

but ${}_\mu\varpi =$ small quantity not zero such that

$$\sum_1^4 {}_\mu\varpi^2 = 0 \quad \text{and} \quad {}_5\varpi = 1. \quad (1.4)$$

Since the direction of ϖ varies for different values of ${}^\mu Q$ ($\mu = 1, 2, 3, 4$) it follows that the four-way measuring system of O does not lie in a fourfold. On the other hand, since the direction of stationary measurements for O (L_5) coincides with 5Q , displacement of the system along 5Q will leave all measurements unchanged. It will be shown in the next section that 5l_4 is related to the 'potential energy' of the field in which O observes P as moving and that it is possible to express the components of ϖ in terms of x^μ , so that the equations of the gravitational field emerge in the required form. This turns on the double limitation on O —his rigid measuring system and his anti-time blindness—the combination of which give the same results as Newton's theory with the small correction introduced by Einstein.

1.8. Electrostatic field

The electrostatic field is defined as the situation which satisfies the conditions ϖ coincident with 5Q and L_5 makes a null angle with 5Q having non-zero components which are functions of iQ .

For this case, in general, displacement of the system along 5Q involves a change in O 's observable co-ordinates x^1, x^2, x^3, x^4 .

The components ${}^1l_5, {}^2l_5, {}^3l_5, {}^4l_5, {}^5l_5$ of the unit vector L_5 in the Q co-ordinate system, i.e. its direction cosines relative to the 5Q , are real and imaginary numbers satisfying

$${}^5l_5 = 1 \quad \text{and} \quad {}^1l_5^2 + {}^2l_5^2 + {}^3l_5^2 + {}^4l_5^2 = 0. \quad (1.5)$$

In the next section it will be shown that 4l_5 is directly related to the electrostatic potential energy observed by O for a P -body carrying unit charge. The angle $\frac{1}{2}\pi - \lambda$ which the cosmodesic of P makes with 5Q is related to the charge E of P ($\tan \lambda \propto E/m_0$). We have thus the necessary elements for constructing the electrostatic field equations. The magnetic vector potential appears when the rotation of the vector set L_μ relative to the nQ is given a velocity rotation relative to the Q and O (i.e. a Minkowski rotation about 5Q) corresponding to the motion of the charged field-producing body M relative to Q and O . Electromagnetic fields thus arise where the four-way measuring system of O lies in a fourfold, but the direction of anti-time for O is not unique.

1.9. *Physical rigidity and geometrical torsions*

Both types of field must be identified with the interpretation of the fact that the measuring system of O is constrained.

The notion of a rigid body is derived from our common experience of persistent material objects. Since we are not concerned with any processes proceeding in the interior of such a body, it is irrelevant to consider its atomic structure. It may, however, be observed that the measuring instruments used in dynamical observations are constructed of materials approximating as closely as possible to ideal rigid bodies, and the assumption that such instruments are available is common to all types of field theory.

The significant properties which we have to discriminate for the purpose of our analysis are those of rigidity and constraint.

It will be shown that the presence of a field is equivalent to the curvatures of the two-dimensional surfaces, traced in the five-dimensional manifold by the ends of a rigid ruler; this curvature arising because, unless the ruler is at right angles to the direction of the field, its two ends are in regions where the angles between the L_k and the kQ are different. It can also be shown that the variation of these angles with the variation of x^μ implies that the L_k forms an axis system with *torsions* (Cartan 1932).

In the gravitational case, the magnitude of the unobservable displacement Δx^5 along the invisible fifth direction corresponding to the interval RS is not uniquely determined by the numbers Δx^μ , but depends, for example, upon whether the interval is traversed by taking Δx^ν or Δx^4 first. In the electromagnetic case the curvature of the envelope of L_5 is not unique but depends upon the ratio of d^5q and d^4q , i.e. upon the angle λ .

2. THE INTERPRETATION OF ABSOLUTE DISPLACEMENTS AS PHYSICAL MEASUREMENTS

2.1. *General field conditions*

In the previous section we have established two independent geometrical constructions both referred to the Q -co-ordinate system of the cosmic manifold. One

is the cosmodesic of the unconstrained point body P , determined by the angle $\frac{1}{2}\pi - \lambda$ which it makes with 5Q . The second is the rigid measuring system of O , determined by the unit vectors L_j each making small, null or zero angles with the corresponding kQ axes through the given point. These constitute the whole equipment required for constructing a general unified field theory.

Let the unit vector L_j have direction cosines kl_j with reference to the orthogonal reference frame kQ such that

$$\sum_{k=1}^5 {}^kl_j {}^kl_j = 1. \quad (2.1)$$

The kl_j are assumed to be functions of nQ . This assumption is in fact the introduction of the most general kind of field possible. In order to obtain results which will be applicable to O 's observations and measurements, it is necessary to specify the conditions under which the kl_j 's may be expressed as functions of the space-time quantities x^μ , where the x^μ are measured along curves in the cosmic manifold to which the L_μ are tangent vectors ($\mu = 1, 2, 3, 4$). This is done in equation (2.2). First, however, we must develop the general rigidity conditions limiting the choice of the kl_j 's.

Consider an infinitesimal element $ds = RS$ of the cosmodesic of P passing through R , the origin of the kQ co-ordinate system. The point S has infinitesimal displacements d^kq from R . Let Δx^j be the infinitesimal displacements along the five directions L_j corresponding to d^kq . By definition of the kl_j

$$d^kq = {}^kl_j \Delta x^j, \quad (2.2)$$

whence

$$\Delta x^n = {}_k l^n d^kq, \quad (2.3)$$

where ${}_k l^n$ is the cofactor of kl_n in $\| {}^kl_j \|$ divided by the value of this determinant.

We now make use of the approximative assumption that the components of the L_j are nearly the same as those of the jQ . Thus if ϵ is a small number, we have, at most

$$\left. \begin{array}{cccccc} {}^1\varpi, & {}^1l_4, & {}^1l_5, & {}^4l_\psi, & {}^5l_\psi & = \sqrt{(-1)} O(\epsilon), \\ {}^4\varpi, & {}^1l_\eta, & {}^5l_4, & {}^4l_5 & & = O(\epsilon), \end{array} \right\} \quad (2.4)$$

whence it follows that

$${}^il_j = 1 + O(\epsilon^2).$$

The orthogonality conditions required by the rigidity of O 's measuring system (cf. § 1.5) can now be stated in differential form.

(i) ϖ is orthogonal to L_μ giving

$${}_1\varpi + {}^5l_1 + {}_2\varpi + {}^2l_1 + {}_3\varpi + {}^3l_1 + {}_4\varpi + {}^4l_1 = iO(\epsilon^3), \quad (2.5)$$

with three similar equations for the kl_2 , kl_3 and kl_4 .

(ii) The L_ψ are mutually orthogonal, giving

$${}^1l_2 + {}^2l_1 + {}^3l_1 {}^3l_2 + {}^4l_1 {}^4l_2 + {}^5l_1 {}^5l_2 = O(\epsilon^3), \quad (2.6)$$

with two similar equations for the pairs (2, 3) and (3, 1).

(iii) L_ψ is orthogonal to the projection of L_4 into ${}^5Q = 0$, giving

$${}^1l_4 + {}^4l_1 + {}^2l_1 {}^2l_4 + {}^3l_1 {}^3l_4 = iO(\epsilon^3), \quad (2.7)$$

with two similar equations for the pairs (2, 4) and (3, 4).

These conditions restrict the choice of the directions L_μ for both types of field. They are the minimum requirements which must be satisfied if O is to make measurements which are self-consistent in time. It can readily be seen that this is physically equivalent to the postulate that O 's measuring system coheres in such a way as to be self-identical in time to first-order accuracy.

2.2. The condition that the kl_j 's can be expressed as functions of x^μ

2.2.1. Gravitational field

Here we have L_5 in the direction of 5Q so that

$${}^5l_5 = 1, \quad {}^\mu l_5 = 0 \quad \text{and hence} \quad {}_5l^\mu = 0. \quad (2.8)$$

It follows that equations (2.3) do not contain d^5q . Hence there is no loss of generality in considering only cosmodesics lying in ${}^5Q = 0$.

Now in the gravitational field the L_μ do not lie in ${}^5Q = 0$, and hence the Δx^μ (when expressed as functions of d^kq) must be integrable for all possible d^kq . The conditions for this are

$$\frac{\partial({}_\nu l^\mu)}{\partial^\sigma q} = \frac{\partial({}_\sigma l^\mu)}{\partial^\nu q} \quad (\nu \neq \sigma). \quad (2.9)$$

Equations (2.9) are thus the general differential equations of condition in the gravitational field. When they are satisfied the kl_j which are defined as functions of nq can be expressed as functions of x^μ .

2.2.2. General formulae for the co-factors ${}_k l^j$

For the purpose of developing (2.10) explicitly it is convenient to obtain the co-factors ${}_k l^j$ evaluated to $O(\epsilon^3)$. We find, using (2.4) and ${}^5l_5 = 1$,

$$\left. \begin{aligned} {}_k l^j &= -3 {}^j l_k + \sum_{n=1}^5 {}^n l_k {}^j l_n + O(\epsilon^3) \quad (j \neq k), \\ {}_k l^k &= 2 - {}^k l_k + \sum_{\substack{n=1 \\ n \neq k}}^5 {}^n l_k {}^k l_n + O(\epsilon^3). \end{aligned} \right\} \quad (2.10)$$

Making use of (2.6) and (2.7) we obtain the alternative expressions

$$\left. \begin{aligned} {}_\xi l^\eta &= {}^\xi l_\eta + {}^5l_\xi ({}^5l_\eta + {}^\eta l_5) + O(\epsilon^3), \\ {}_4 l^\eta &= {}^4 l_\eta + {}^5l_4 {}^\eta l_5 + O(\epsilon^3), \\ {}_\eta l^4 &= {}^\eta l_4 + {}^5l_\eta {}^4 l_5 + O(\epsilon^3). \end{aligned} \right\} \quad (2.11)$$

Equation (2.5) can also be put in the form that ${}_\mu \varpi$ equals ${}_\mu l^5$, a result which will be used later.

2.2.3. *The co-factors in the gravitational field*

For the co-factors, from (2.8), (2.10) and (2.11) we have

$$\left. \begin{aligned} \xi l^\eta &= \xi l_\eta + {}^5l_\xi {}^5l_\eta + O(\epsilon^3) \quad (\xi \neq \eta), \\ {}^4l^\eta &= {}^4l_\eta + O(\epsilon^3), \\ {}_\eta l^4 &= {}_\eta l_4 + O(\epsilon^3), \\ {}_\mu l^5 &= -{}^5l_\mu + \sum_{\substack{n=1 \\ n \neq \mu}}^4 {}^n l_\mu {}^5l_n + O(\epsilon^3), \\ {}_5 l^\mu &= 0, \\ {}_5 l^5 &= 1. \end{aligned} \right\} \quad (2.12)$$

 2.2.4. *Electromagnetic field*

Here ϖ is in the direction of 5Q so that

$${}_5 l^5 = 1, \quad {}_\mu l^5 = 0, \quad (2.13)$$

${}_5 l_5$ being unity as before. In the electromagnetic field the L_μ necessarily lie in ${}^5Q = 0$, so that displacements along 5Q have no meaning for O . Hence it is not necessary that all the Δx^μ shall be integrable with respect to such displacements in order that the ${}^k l_j$ shall be expressible in terms of O 's co-ordinates x^μ . It is, however, necessary that they shall all be integrable with respect to displacements $d^\mu q$; and also that Δx^4 shall be integrable with respect to displacements $d^5 q$ as well as displacements $d^\mu q$, for otherwise a charged body situated at the same place as O and having instantaneously zero velocity would disagree with O about the simultaneity of arrival of signals. We have therefore

$$\frac{\partial_\nu l^\xi}{\partial^\mu q} = \frac{\partial_\mu l^\xi}{\partial^\nu q} \quad \text{and} \quad \frac{\partial_k l^4}{\partial^m q} = \frac{\partial_m l^4}{\partial^k q}. \quad (2.14)$$

 2.2.5. *Co-factors in the electromagnetic field*

For the co-factors from (2.11), (2.13) we have

$$\left. \begin{aligned} {}_\mu l^\nu &= {}^\mu l_\nu + O(\epsilon^3), \\ {}_5 l^\mu &= {}^\mu l_5 + \sum_{\substack{n=1 \\ n \neq \mu}}^4 {}^n l_5 {}^\mu l_n + O(\epsilon^3), \\ {}_\mu l^5 &= 0, \\ {}_5 l^5 &= 1. \end{aligned} \right\} \quad (2.15)$$

The symmetry of the two kinds of field can be seen at once by comparing (2.12) and (2.15). It remains to show that they have the correct properties to entitle them to be interpreted as gravitational and electromagnetic fields respectively.

3. POTENTIAL ENERGY AND THE GENERALIZED LAGRANGIAN

 3.1. *The relation between constrained and free systems*

Our task is to translate out of the universal kQ co-ordinate system the relations between the two sets O and P in such a way as to express them in terms of O 's physical measurements. This must be done without setting up a co-ordinate system

for O , because, as we have seen in (1.7), O 's measuring system need not lie in a four-fold. Moreover, such a procedure would leave us without the means of allowing for O 's 'anti-time' blindness. We must therefore confine ourselves to the system of directions along which O makes his measurements at a point and then find the means of integrating the differential equations which express the fact that O uses a rigid space-extended structure.

If we fix our attention on the vector L_4 along which O measures time with his clocks, we have a space-distributed variation in the components of the null angles between $L_5 - {}^5Q$ and $\varpi - {}^5Q$ respectively. These components ${}^\mu l_5$ and ${}_\mu \varpi$ being variable functions of the x^μ it must follow that the L_4 vectors through different points are not parallel to one another. It follows that the L_4 's make a varying angle with the cosmodesic of P , and this in turn leads to O 's observation of P 's motion as curvilinear and accelerated.

In order to use this result in the analysis, we have to express the variational condition for the straightness of the cosmodesic

$$\delta \int \sqrt{\left\{ \sum_{k=1}^5 (d^k q)^2 \right\}} = 0 \quad (3.1)$$

in terms of O 's physical measurements, i.e. x^1 , x^2 , x^3 and x^4 .

In defining O 's measuring system in § 2 we included an anti-time-like measurement Δx^5 (being the displacement along L_5 measured on the same scale and in homogeneous magnitudes with Δx^μ). This is permissible in the differential form, but Δx^5 must be eliminated before integration, since the proof of § 2 that the Δx^μ are integrable (i.e. have a single value independent of the path of measurement) does not apply to Δx^5 .

$$\text{Now} \quad ds^2 = \sum_{k=1}^5 (d^k q)^2 = \sum_{k=1}^5 ({}^k l_j \Delta x^j)^2$$

from (2.2). In order to eliminate Δx^5 , we use the relation

$${}^5 l_j \Delta x^j = d^5 q = ds \sin \lambda,$$

whence, since as shown in (1.3) and (1.5) ${}^5 l_5$ is always unity,

$$\Delta x^5 = ds \sin \lambda - {}^5 l_\mu \Delta x^\mu. \quad (3.2)$$

$$\text{Thus} \quad ds^2 = \sum_{\mu=1}^5 \{ {}^k l_\mu \Delta x^\mu + {}^k l_5 (ds \sin \lambda - {}^5 l_\mu \Delta x^\mu) \}^2,$$

whence, using (2.4) to obtain approximate values permitted by the postulated smallness of the l 's, and ${}_\mu \varpi$, we obtain the relation*

$$ds^2 \cos^2 \lambda - 2ds \sin \lambda \left\{ \left(\sum_{\nu=1}^4 {}^\nu l_5 {}^\nu l_\mu \right) \Delta x^\mu \right\} = \left\{ {}^5 l_\mu {}^5 l_\nu - 2 {}^5 l_\nu \sum_{k=1}^5 {}^k l_5 {}^k l_\mu + \sum_{k=1}^5 {}^k l_\mu {}^k l_\nu \right\} \Delta x^\mu \Delta x^\nu \quad (3.3)$$

* This equation contains the same terms as that assumed by Mosharrafa (1948) as the metric equation for a 'meta-Riemannian' space, and his proof that it gives rise to terms which are interpreted as gravitational and magnetic vector potentials can therefore be applied here. In the present paper, however, the equation arises naturally as a consequence of the basic assumptions of § 1.2.

between ds and Δx^μ , in which the coefficient of Δx^μ is correct to $O(\epsilon^3)$ and that of $\Delta x^\mu \Delta x^\nu$ to $O(\epsilon^4)$.

This is the quadratic equation for the interval ds along the cosmodesic of P as it is found in O 's measuring system.

3.2. The Lagrangian

In order to exhibit the connexion between the variational equation for the straightness of the cosmodesic and the equations of motion we can define quantities L , v^μ , V^μ and V by the equations

$$\left. \begin{aligned} 1 - \frac{L}{m_0 c^2} &= \frac{ds \cos \lambda}{\Delta x^4}, \\ \frac{\Delta x^\psi}{\Delta x^4} &= \frac{i V^\psi}{c} = v^\psi, \\ \frac{V^4}{c} &= v^4 = 1, \\ \sum_{\psi=1}^3 (v^\psi)^2 &= V^2. \end{aligned} \right\} \quad (3.4)$$

Inserting from these in (3.1) we obtain the form

$$\delta \int \left(1 - \frac{L}{m_0 c^2} \right) \sec \lambda \Delta x^4 = 0. \quad (3.5)$$

Since λ and $m_0 c^2$ are time and space constant, this is equivalent to the classical variational equation $\delta \int L dt = 0$ for the motion of a particle in a field of force.

We can accordingly derive the potential energy of the motion from (3.3) by the following quadratic in L :

$$\left(1 - \frac{L}{m_0 c^2} \right)^2 - 2 \left(1 - \frac{L}{m_0 c^2} \right) \left\{ \left(\sum_{\nu=1}^4 v_\nu^\nu l_\mu^\nu \right) v^\mu \right\} \tan \lambda = \left\{ {}^5 l_\mu {}^5 l_\nu - 2 {}^5 l_\nu \sum_{k=1}^5 k l_5^k k l_\mu^k + \sum_{k=1}^5 k l_\mu^k k l_\nu^k \right\} v^\mu v^\nu. \quad (3.6)$$

In order to interpret this equation in terms of O 's physical measurements, we may consider the approximate forms which are appropriate for the various types of field.

3.3. The gravitational field

Applying the equations of § 2.2 and the approximations permitted by (2.4) to equation (3.3) we have

$$ds^2 \cos^2 \lambda = \sum_{\mu=1}^4 \{ 1 - ({}^5 l_\mu^\mu)^2 \} (\Delta x^\mu)^2 + \left\{ \sum_{\sigma=1}^4 ({}^\sigma l_\mu^\sigma {}^\sigma l_\nu^\sigma) \right\} \Delta x^\mu \Delta x^\nu \quad (\mu \neq \nu).$$

Using (2.6) and (2.7) to evaluate the coefficients of the last term, this gives

$$ds^2 \cos^2 \lambda = \sum_{\mu=1}^4 \{ 1 - ({}^5 l_\mu^\mu)^2 \} (\Delta x^\mu)^2 - {}^5 l_\psi {}^5 l_\eta \Delta x^\psi \Delta x^\eta \quad (\psi \neq \eta), \quad (3.7)$$

which reduces to

$$ds^2 \cos^2 \lambda = \sum_{\mu=1}^4 (\Delta x^\mu)^2 - ({}^5 l_4^4)^2 (\Delta x^4)^2 - \{ {}^5 l_\psi \Delta x^\psi \}^2. \quad (3.7)$$

Where 5l_4 is not zero, and we are dealing with the case of spherical symmetry about a point distant r from R , this equation can be transformed into spherical polar co-ordinates by writing

$$\left. \begin{aligned} ie_\psi &= \frac{{}^5l_\psi}{{}^5l_4}, \\ ir &= e_\psi x^\psi \end{aligned} \right\} \quad (3.8)$$

(this is legitimate since $\sum_{\mu=1}^4 ({}^5l_\mu)^2 \neq 0$, so that $\sum_{\psi=1}^3 (e_\psi)^2 = 1$ and hence e_ψ may be regarded as a set of direction cosines in space at the point R , independent of r)

and

$$dr^2 + r^2 d\theta^2 + r^2 \sin^2 \theta d\phi^2 = - \sum_{\psi=1}^3 (\Delta x^\psi)^2.$$

Then (3.7) takes the form

$$ds^2 \cos^2 \lambda = \{1 - ({}^5l_4)^2\} (\Delta x^4)^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2 - dr^2 \{1 + ({}^5l_4)^2\}. \quad (3.9)$$

Equation (3.9) is recognized immediately as having the same form as the Schwarzschild equation providing the function ${}^5l_4 = f(r, \theta, \phi)$ is correctly assigned. It must again be emphasized that (3.9) is not a metric equation in the sense of determining the space constants of a metrical manifold. It is simply the statement of what O will find when he makes measurements in his own rigidly constructed system upon the motion of the unconstrained body P . It follows that the apparent track of P as observed by O satisfies all the physical requirements of a body falling in a gravitational field including the Einstein correction to the Newtonian theory. In the next section we shall derive the function 5l_4 from the differential equations of condition obtained in § 2.

It will facilitate the work of physical interpretation if we show how the potential energy is derived from (3.6). When we omit the Einstein correction terms (corresponding to $dr^2 ({}^5l_4)^2$ in (3.9)), (3.6) becomes

$$\left(1 - \frac{L}{m_0 c^2}\right)^2 \doteq 1 - ({}^5l_4)^2 - \frac{V^2}{c^2}.$$

Thus

$$L = m_0 c^2 \left\{ 1 - \sqrt{1 - ({}^5l_4)^2 - \frac{V^2}{c^2}} \right\},$$

which may be compared with the classical Lagrangian

$$L = m_0 c^2 \left\{ 1 - \sqrt{1 - \frac{V^2}{c^2}} \right\} - \Omega_g;$$

clearly the potential energy Ω_g of any body in a field of this type is proportional to its inertial rest-mass, and is always negative (giving an attractive force). This follows from the fact that 5l_4 is a real number and we have approximately

$$\Omega_g \doteq -m_0 c^2 \cdot \frac{1}{2} ({}^5l_4)^2. \quad (3.10)$$

3.4. The electromagnetic field

Inserting ${}^5l_\mu = 0$ and making use of the approximation of (2.4) we get (3.6) in the form

$$\left(1 - \frac{L}{m_0 c^2}\right)^2 - 2\left(1 - \frac{L}{m_0 c^2}\right) \left\{ \sum_{\nu=1}^4 {}^\nu l_5 {}^\nu l_\mu v^\mu \right\} \tan \lambda = \left\{ \sum_{\sigma=1}^4 {}^\sigma l_\mu {}^\sigma l_\nu \right\} v^\mu v^\nu = \sum_{\mu=1}^4 (v^\mu)^2 + O(\epsilon^3)$$

by (2.6) and (2.7).

Now we can make use of the fact that by hypothesis the potential and kinetic energies are both small compared with unity and complete the square on the left-hand side to obtain

$$L \doteq m_0 c^2 \left[1 - \sqrt{\left(1 - \frac{V^2}{c^2}\right)} - \left(\sum_{\nu=1}^4 {}^\nu l_5 {}^\nu l_\mu \right) v^\mu \tan \lambda \right].$$

In this expression—to exhibit a familiar form—we have retained terms under the radical of the same order as those neglected in the approximations.

We have thus a field of force with potential energy Ω_e given by

$$\Omega_e = m_0 c^2 \tan \lambda \left(\sum_{\nu=1}^4 {}^\nu l_5 {}^\nu l_\mu \right) v^\mu. \quad (3.11)$$

Now λ is a property of the unconstrained body P and is independent of the field (being simply the angle by which its cosmodesic diverges from perpendicularity to 5Q). We can therefore write

$$m_0 c^2 \tan \lambda = \frac{E}{\zeta}, \quad (3.12)$$

where ζ is an universal constant and E is the electric charge upon P .

From this it can be seen that we are dealing with a field which does not act on bodies in direct proportion to their inertial mass. In particular, a body whose cosmodesic lies in ${}^5Q = 0$ has $\lambda = 0$ and therefore zero potential energy everywhere. Moreover, fields of this type can be attractive or repulsive according to the signs of the numbers $\tan \lambda$ and ${}^\nu l_j$, one of which belongs to P itself and the others to the field.

In order to complete the demonstration that we have all the properties of electric fields, we need to show how to express Ω_e/E in terms of a vector potential. Since, according to the basic assumption for this type of field, L_1, L_2, L_3 and L_4 all lie in ${}^5Q = 0$ and are mutually orthogonal, the ${}^\nu l_\mu$ can only correspond to the direction cosines of a Minkowski velocity rotation. Referring the latter to the kQ system, we can assign real components (${}^1U, {}^2U, {}^3U$) to the velocity and define the mixed quantities ${}^1u, {}^2u, {}^3u, {}^4u$ and u by the equations

$$\frac{i\psi U}{c} = \psi u, \quad {}^4u = 1, \quad u^2 = ({}^1u)^2 + ({}^2u)^2 + ({}^3u)^2. \quad (3.13)$$

$$\left. \begin{aligned} \text{Then the } {}^\nu l_\mu \text{ are given by } \quad & \psi l_\psi = 1 - \frac{(\psi u)^2}{2}, \\ & \psi l_\eta = -\frac{\psi u \eta u}{2} = \eta l_\psi, \quad (\psi \neq \eta) \\ & {}^4 l_\psi = \psi u = -\psi l_4, \\ & {}^4 l_4 = 1 - \frac{u^2}{2}, \end{aligned} \right\} \quad (3.14)$$

substituting these values in (3.11), assuming that ψu and u are $O(\epsilon)$ and omitting terms in $O(\epsilon^3)$, we obtain

$$\frac{\zeta \Omega_e}{E} = \sum_{\psi=1}^3 \psi \{ \psi l_5 + \psi u^4 l_5 \} + \left\{ \sum_{\psi=1}^3 \psi l_5 \psi u + {}^4 l_5 \right\}. \quad (3.15)$$

This differs from the classical form for the potential energy of a charged body moving in a field which has magnetic four-vector potential

$$\mathbf{A} \equiv {}^4 l_5 ({}^1 u, {}^2 u, {}^3 u, 1), \quad (3.16)$$

only by the presence of terms in ψl_5 . Of these, the one entering the electrostatic component, viz. $\sum_{\psi=1}^3 \psi l_5 \psi u$, is a small correction on the main part ${}^4 l_5$ of this component because ψl_5 is $O(\epsilon)$ and so is ψu .

The terms in ψl_5 in the magnetic components appear to be large compared with the classical term $\psi u^4 l_5$, but they do not give rise to any forces because the magnetic force is given by $\text{curl } \mathbf{A}$ and $\text{curl } ({}^4 l_5)$ vanishes identically. This follows because

$$\sum_{\nu=1}^4 (\nu l_5)^2 = 0, \text{ so that } \psi l_5 \text{ can be expressed as } f_{\psi} {}^4 l_5 \text{ where } \sum_{\psi=1}^3 (f_{\psi})^2 = 1. \quad (3.17)$$

Hence ψl_5 are the components of a space-vector of magnitude ${}^4 l_5$ and direction $\mathbf{n} \equiv (f_1 f_2 f_3)$. From considerations of spatial symmetry it can be seen that this direction \mathbf{n} coincides with the direction of maximum rate of change of ${}^4 l_5$. It follows that ψl_5 are the components of the gradient of the scalar $\int {}^4 l_5 dn$, and hence $\text{curl } (\psi l_5)$ vanishes identically. The term $\text{curl } \psi u^4 l_5$ does not of course vanish in this way because the vector \mathbf{u} is not parallel to \mathbf{n} . The demonstration of the electromagnetic field is thus complete.

4. THE INVERSE SQUARE LAW OF FORCE

4.1. The conservation principle

The expressions we have derived for potential energy have been shown to possess the properties appropriate for the gravitational and electromagnetic potential at a point. It is necessary to show also that these are distributed in space according to the inverse square law of force. We shall here undertake this demonstration to the same approximation as we obtained for the co-factor in § 2.

It is easy to see that the inverse square law is associated with the time constancy of the field and the three-dimensionality of space. In view of the distinction we have drawn between Q and O , we cannot make an assumption of the type used in classical theory, as to the vanishing of the divergence of a gradient, since this is no longer obviously an 'absolute' property of O 's material objects in the cosmic manifold. We have, however, in the physical conservation principle, geometrically interpreted according to § 1.2 a property which must apply to both types of field.

The physical principle of conservation of energy is connected with the condition of integrability of the measurements of O . These conditions were given in § 2, equations (2.9) and (2.14). We have to demonstrate how these conditions lead to

the inverse square law for precisely the two kinds of potential energy we have found and no others that are not merely combinations or derivations of these. We do not use all of equations (2.9) and (2.14), nor shall we work, in the present paper, to a higher degree of approximation than in § 2.

4.1.2. Relative rigidity

A physical observer O measures space with a ruler, which we can represent by a box having sides of length $\Delta X_1, \Delta X_2, \Delta X_3$ along the vectors L_1, L_2, L_3 . It will be recalled that these three vectors are mutually orthogonal, so the volume of the box is $\Delta X_1 \cdot \Delta X_2 \cdot \Delta X_3$. Let A, B be the vertices of one of the sides, say ΔX_3 , of the box. Translate A to A' along L_4 , i.e. in the direction ${}^k l_4$ through a displacement δx^4 . Similarly, translate B to B' along L'_4 , the time-line through B having direction ${}^k l'_4$ through a displacement δx^4 . To the first order

$${}^k l'_4 = {}^k l_4 + \frac{\partial {}^k l_4}{\partial x^3} \Delta X_3. \quad (4.1)$$

But in Q 's measurement, using (2.2), and taking A as the origin of co-ordinates,

$$\left. \begin{aligned} \text{co-ordinates of } B &= {}^k l_3 \cdot \Delta X_3, \\ \text{co-ordinates of } A' &= {}^k l_4 \cdot \delta x^4, \\ \text{co-ordinates of } B' &= {}^k l_3 \cdot \Delta X_3 + {}^k l'_4 \cdot \delta x^4, \end{aligned} \right\} \quad (4.2)$$

whence

$$(A'B')^2 = (\Delta X_3)^2 \left[\sum_{k=1}^5 \left({}^k l_3 + \frac{\partial {}^k l_3}{\partial x^3} \delta x^4 \right)^2 \right] = (\Delta X_3)^2 \left[1 + 2 \sum_{k=1}^5 {}^k l_3 \frac{\partial {}^k l_3}{\partial x^3} \delta x^4 + O(\delta x^4)^2 \right],$$

$$\text{so that to } O(\delta x^4), \quad A'B' = \Delta X_3 \left[1 + \frac{\partial {}^3 l_4}{\partial x^3} \delta x^4 + O\{(\delta x^4)^2, \epsilon \epsilon_1\} \right], \quad (4.3)$$

$$\text{where} \quad \epsilon_1 = O\left(\frac{\partial \epsilon}{\partial x^\psi}\right) = O\left(\frac{\partial \epsilon}{\partial x^q}\right). \quad (4.4)$$

Evidently the box remains, to the first order, a box, and its volume changes by

$$(\Delta X_1 \Delta X_2 \Delta X_3) (\delta x^4) \left(\sum_{\eta=1}^3 \frac{\partial \eta l_4}{\partial x^\eta} \right), \quad (4.5)$$

where the term $O(\delta x^4 \epsilon \epsilon_1)$ has been neglected. It follows that the volume of the box remains constant for Q if

$$\sum_{\eta=1}^3 \frac{\partial \eta l_4}{\partial x^\eta} = 0. \quad (4.6)$$

4.2. The gravitational potential

To derive the law of force we make use of the condition (2.9) that x^4 shall be integrable, namely,

$$\frac{\partial ({}_4 l^4)}{\partial \eta q} = \frac{\partial ({}_\eta l^4)}{\partial {}^4 q} \quad (\eta = 1, 2, 3). \quad (4.7)$$

$$\text{From (2.7, 8, 10)} \quad {}_4 l^4 = 1 + \frac{1}{2} ({}^5 l_4)^2 - \frac{1}{2} \Psi^2 + O(\epsilon^3), \quad (4.8)$$

$$\text{where} \quad \Psi^2 = \sum_1^3 ({}^\eta l_4)^2, \quad (4.9)$$

and Ψ is the velocity (for O) of the freely falling body having cosmodesic Q^4 .

Also from (2.7)

$$\eta l^4 = \eta l_4 + O(\epsilon^3). \quad (4.10)$$

Hence (4.7) becomes

$$\frac{1}{2} \frac{\partial}{\partial \eta q} \{({}^5l_4)^2 - \Psi^2\} = \frac{\partial}{\partial^4 q} \{\eta l_4\} + O(\epsilon^2 \epsilon_1, \epsilon^2 \epsilon'), \quad (4.11)$$

where*

$$\epsilon_s = O\left(\frac{\partial^s \epsilon}{\partial \eta q^s}\right), \quad \epsilon' = O\left(\frac{\partial \epsilon}{\partial^4 q}\right). \quad (4.12)$$

Now

$$\frac{\partial}{\partial \eta q} = \sum_{k=1}^5 \frac{\partial x^k}{\partial \eta q} \frac{\partial}{\partial x^k} = \sum_{k=1}^4 \eta l^k \frac{\partial}{\partial x^k},$$

whence

$$\frac{\partial}{\partial \eta q} = \frac{\partial}{\partial x^\eta} \{1 + O(\epsilon^2)\} + O\left(\epsilon \frac{\partial}{\partial x^k}\right) \quad (k \neq \eta). \quad (4.13)$$

Hence (4.11) may be written

$$\frac{\partial}{\partial x^\eta} \{({}^5l_4)^2 - \Psi^2\} = \frac{\partial}{\partial^4 q} (\eta l_4) + O(\epsilon^2 \epsilon_1, \epsilon^2 \epsilon'). \quad (4.14)$$

Differentiate (4.14) with respect to x^η and sum over $\eta = 1 \dots 3$, obtaining,

$$\nabla^2 \{({}^5l_4)^2 - \Psi^2\} = 2 \frac{\partial}{\partial^4 q} \left(\sum_{\eta=1}^3 \frac{\partial \eta l_4}{\partial x^\eta} \right) + O(\epsilon \epsilon_1^2, \epsilon \epsilon_1 \epsilon', \epsilon^2 \epsilon_2, \epsilon^2 \epsilon'), \quad (4.15)$$

where

$$\nabla^2 = \sum_{\eta=1}^3 \frac{\partial^2}{\partial (x^\eta)^2}. \quad (4.16)$$

Now for O 's time x^4 , $\sum_{\eta=1}^3 \frac{\partial \eta l_4}{\partial x^\eta}$ vanishes in virtue of the assumption made in (4.1), and

we have, to the present order of approximation,†

$$\nabla^2 \{({}^5l_4)^2 - \Psi^2\} = O(\epsilon^2 \epsilon_2, \epsilon \epsilon_1^2). \quad (4.17)$$

It was shown in § 3 that $\frac{1}{2}({}^5l_4)^2$ is the gravitational potential and clearly the Lagrangian implies, if relativistic corrections are neglected as is the case in the present section, conservation of energy in the form

$$\frac{1}{2}({}^5l_4)^2 + \frac{1}{2}\Psi^2 = \text{constant}. \quad (4.18)$$

whence (4.17) yields

$$\nabla^2 \left\{ \frac{1}{2}({}^5l_4)^2 \right\} = 0. \quad (4.19)$$

So that the gravitational potential satisfies Laplace's equation.

4.3. The electrostatic potential

From the second set of equations in (2.14), one of the conditions of integrability of x^4 is

$$\frac{\partial({}_\eta l^4)}{\partial^5 q} = \frac{\partial({}_5 l^4)}{\partial \eta q} = -\frac{\partial({}_4 l^5)}{\partial \eta q} + O(\epsilon \epsilon_1), \quad (4.20)$$

or

$$\frac{\partial({}_\eta l^4)}{\partial^5 q} = -\frac{\partial({}_4 l^5)}{\partial x^\eta} + O(\epsilon \epsilon_1, \epsilon \epsilon'), \quad (4.21)$$

* In a more precise treatment to be given later, it will be shown that the integrability conditions imply that not all of the ${}_\eta l^k$ ($k \neq j$) can be $O(\epsilon)$; some must be of a greater order of small quantities.

† I.e. taking $O(\epsilon') = O(\epsilon_1)$.

by the same steps as for the gravitational potential. Hence

$$\nabla^2({}_4I^5) = \frac{\partial}{\partial^5 q} \left\{ \sum_{\eta=1}^3 \frac{\partial^{\eta} I^5}{\partial x^{\eta}} \right\} + O(\epsilon\epsilon_2, \epsilon_1^2), \quad (4.22)$$

giving Laplace's equation approximately as before.

It will be noted that both the equations governing the law of force follow from the condition of relative rigidity given in § 4.1—in a later publication a more exact treatment will be given. For the present purpose it is sufficient to record that the Laplace equation is evidently a consequence of the condition that x^4 shall be a curve in the cosmic manifold uniquely determined at every point of a rigid body.

5. THE GRAVITO-MAGNETIC FIELD

5.1. *The field of a rotating massive uncharged rigid body*

As discussed in § 1, the concept of rigidity is as fundamental to the present theory as it is to classical dynamics. We have, moreover, implied that absolute rotations exist by postulating a Q system of co-ordinates with respect to which such rotations can be measured. Now consider the case in which a relatively light mass δM rotates around a heavy mass M in a circular path. If δM is freely falling, i.e. held to M only by gravitational or electromagnetic attraction, it describes a cosmodesic, and its own field will be that produced by a body of its charge and mass rotating as it rotates—the equations of §§ 3 and 4 being sufficient to give this field. Let u (defined as in § 3.4, equations (3.13)) represent the Minkowski rotation in the Q system corresponding to the velocity of δM . Then this means that at any given moment the displacement of O 's $L_1 \dots L_5$ from ${}^1Q \dots {}^5Q$ at another point R will be that corresponding to the charge and mass of δM , but with the Minkowski rotation about 5Q corresponding to the velocity of δM .

Suppose now that δM is connected by a rigid arm to M , but still rotates at such a velocity that centrifugal force balances gravitational attraction. At first sight the situation would appear to be unchanged, since the motion of δM is kinematically identical, and there is no force in the arm. But clearly the potentialities of the situation are now quite different; for example, δM would respond quite differently to a frictional retardation or an impact with another body. Hence the situation with regard to anti-time must be different. The difference is, in fact, that instead of the motion of δM producing the difference between $L_1 \dots L_5$ and ${}^1Q \dots {}^5Q$ corresponding to the rotation u about 5Q , this difference is rotated by u about the line orthogonal to $L_1 \dots L_4$, viz. ω . This effect will of course apply equally whether u is the velocity which δM would have if it were unattached or any other velocity, but it is of special importance to show that it does not vanish even in the former case.

We shall now show that, in the case of the rigid attachment of a rotating, uncharged but massive δM to M , δM produces a magnetic field as if it had a charge numerically equal to its gravitational mass but without the corresponding electric field.

We define the following sets of co-ordinates and directions. L_{μ} and x^{μ} are respectively the unit vectors and the measurements of the physical observer O at the

point R , made under the influence of the constrained motion of δM . ω is orthogonal to the L^μ .

L'_μ and x'^μ are the unit vectors and the measurements which O would make at the point R , if δM were present but had no constrained motion. Hence

$$d^k q = {}^k l'_j \Delta x'_\mu, \quad (5.1)$$

where the ${}^k l'_j$ are written for the gravitational ${}^k l_j$ of (2.2) for the field of δM . ω' , defined as the direction through R orthogonal to L'_μ , coincides with ω because the velocity of the field is constrained.

Since L_μ and L'_μ are not orthogonal sets of vectors, it is not possible to express the relation between them simply by a Minkowski rotation of the form given in equation (3.14). We therefore use orthogonal co-ordinate sets Y^k , Y'^k , which are related to one another by the Minkowski rotation u .

The direction of Y^ψ for $\psi = 1, 2, 3$ is that of L_ψ , whilst Y^5 is in the direction of ω . Y^4 is then orthogonal to all the Y^ψ and to Y^5 . The Y'^k are similarly related to L'_ψ and ω' . Now Y^5 , Y'^5 have the same direction ω , so that

$$dY'^\mu = {}^\mu \lambda_\nu dY^\nu, \quad (5.2)$$

where

$$\left. \begin{aligned} {}^\psi \lambda_\psi &= 1 - \frac{(u^\psi)^2}{2}, \\ {}^\psi \lambda_\eta &= -\frac{u^\psi u^\eta}{2} = +{}^\eta \lambda_\psi, \\ {}^4 \lambda_\psi &= u^\psi = -{}^\psi \lambda_4, \\ {}^4 \lambda_4 &= 1 - \frac{u^2}{2}. \end{aligned} \right\} \quad (5.3)$$

Writing

$$d^k q = {}^k \mu'_j dY'^j, \quad (5.4)$$

we have from equations (5.1) and (2.5)

$$\left. \begin{aligned} {}^k \mu'_\psi &= {}^k l'_\psi \quad (\psi = 1, 2, 3), \\ {}^k \mu'_5 &= {}^k \omega, \end{aligned} \right\} \quad (5.5)$$

and from the orthogonality of Y'^4 to Y'^ψ , Y'^5

$$\left. \begin{aligned} \sum_1^5 ({}^k \mu'_4)^2 &= 1, \\ \sum_1^5 ({}^k \mu'_\psi {}^k \mu'_4) &= 0, \\ \sum_1^5 ({}_p \omega^p {}^k \mu'_4) &= 0. \end{aligned} \right\} \quad (5.6)$$

From (5.2) and (5.4)

$$d^k q = ({}^k \mu'_j {}^\lambda \lambda_j) dY^s, \quad (5.7)$$

where we have taken

$${}^\mu \lambda_5 = {}^5 \lambda_\mu = 0, \quad {}^5 \lambda_5 = 1. \quad (5.8)$$

Now by definition (equation (2.3))

$$\Delta x'^v = {}_n l'^v d^n v = ({}_k l'^v {}^k \mu'_s) dY'^s,$$

and by the construction given above, the corresponding relation between Δx^v and dY^s is

$$\Delta x^v = ({}_k l'^v{}_k \mu'_s) dY^s = {}_k l^v d^k q \quad (\text{by definition of } {}^k l_v), \quad (5.9)$$

using (5.7) we must have ${}_k l^v{}_k \mu'_j \lambda_s = {}_k l'^v{}_k \mu'_s$,

which by way of

$${}_k l_{\omega}{}_k l^v = \delta_{\omega}^v$$

leads to

$${}_p l_{\omega} = {}_p \mu'_j \lambda_s {}_k l \mu'^s{}_k l'_{\omega}. \quad (5.10)$$

Further, it follows directly from the construction that

$${}_p l_{\omega} {}_p l_v = {}_p l'_{\omega} {}_p l'_v. \quad (5.11)$$

If now we return to the general expression for the Lagrangian (3.6) we can see from the foregoing that the cross-product terms in $\Delta x^{\mu} \Delta x^{\nu}$ are the same as those obtained in (3.7) for the gravitational field. There is, however, a linear term in Δx^{μ} because ${}^{\nu} l_5 \neq {}^{\nu} l'_5$ and the former therefore are not zero; whence it follows that

$$\sum_{\nu=1}^4 {}^{\nu} l_5 {}^{\nu} l_{\mu} \neq 0. \quad (5.12)$$

We may choose ${}^{\nu} l_{\mu}$ to be of the form

$${}^{\nu} l_{\nu} = 1 + O(\epsilon^2), \quad {}^{\nu} l_{\mu} = O(\epsilon) \quad (\nu \neq \mu).$$

It is necessary, however, to calculate ${}^{\nu} l_5$. From equations (5.10) and (2.8)

$${}_p l_5 = {}_p \mu'_j \lambda_s {}_5 l \mu'^s. \quad (5.13)$$

It is easy to evaluate ${}^{\psi} l_5$ if terms in $(u^{\psi} \cdot u^{\eta})$ are neglected; for by (5.3) and (5.8)

$${}^{\psi} l_5 = {}^{\psi} \mu'_j {}_5 l \mu'^j - {}^{\psi} \mu'_{\eta} u^{\eta} {}_5 l \mu'^4 + {}^{\psi} \mu'_4 u^{\eta} {}_5 l \mu'^{\eta},$$

where we have used the fact that ${}^i \lambda_j = 1 + O(u^2)$ in deriving the first term on the right-hand side. But, by equation (2.10),

$${}_5 l \mu'^{\nu} = -{}^{\nu} \mu'_5 + \sum_{\substack{n=1 \\ n \neq \nu}}^4 ({}^n \mu'_5 {}^{\nu} \mu'_n) + O(\epsilon^3) = -{}^{\nu} \varpi + \sum_{\substack{n=1 \\ n \neq \nu}}^4 ({}^n \varpi {}^{\nu} \mu'_n) + O(\epsilon^3),$$

by equation (5.5). Then since ${}^{\psi} \mu'_j {}_5 l \mu'^j = 0$ and neglecting terms of order $(\epsilon^3, u\epsilon^2)$

$${}^{\psi} l_5 = {}^{\psi} \mu'_{\eta} u^{\eta} {}_4 \varpi - {}^{\psi} \mu'_4 u^{\eta} {}_{\eta} \varpi$$

or

$${}^{\psi} l_5 = u^{\psi} {}_4 \varpi + O(\epsilon^2 u, u^2, \epsilon^3). \quad (5.14)$$

Finally,

$${}^4 l_5 = {}^4 \mu'_j {}_5 l \mu'^j - {}^4 \mu'_{\eta} u^{\eta} {}_5 l \mu'^4 = {}^4 \mu'_{\eta} u^{\eta} {}_4 \varpi = ({}^4 l'_{\eta} u^{\eta}) {}_4 \varpi,$$

or

$${}^4 l_5 = O(\epsilon^3, u^2). \quad (5.15)$$

By a similar procedure we find that ${}^5 l_5$ is equal to unity to the order (ϵ^3, u^2) . Substituting these values in equations (3.11) and neglecting $u^{\psi} {}^4 l_5$ in comparison with ${}^{\psi} l_5$ as (5.14) and (5.15) have shown to be permissible, we find a potential energy Ω_e for a body having charge E and moving with velocity v in the field of δM given by

$$\frac{\zeta \Omega_e}{E} = \sum_{\psi=1}^3 {}_4 \varpi \{u^{\psi} v^{\psi}\} + O(\epsilon^3, u^2). \quad (5.16)$$

Comparison of (5.16) with (3.15) shows that the first three terms correspond to an electromagnetic field of magnitude such that ${}_4\omega$ replaces ${}_4l_5$; now by (2.5) ${}_4\omega$ is the same as $-{}^5l_4$ which is the gravitational rotation due to δM ; from this it follows that a moving gravitational unit of matter δM which attracts an equal mass at 1 cm. distance with a force of one dyne, will produce a magnetic field corresponding to an electrostatic unit of charge defined the same way and having the same velocity. This is Wilson's (1923) hypothesis to account for the magnetic field of the earth. The fourth or electrostatic term is not ${}_4\omega$ being an order of magnitude in u^7 smaller. It follows that the gravitomagnetic field should not exhibit an observable electrostatic part.

In the case of a rotating rigid massive sphere, integration of this magnetic field gives a magnetic dipole of moment P given by Blackett's (1947) relation

$$P = \beta \frac{G^{\frac{1}{2}}}{c} U, \quad (5.17)$$

where U is the angular momentum and $\beta = 1$. It will be noted that the derivation requires that the rotating body should be rigid and that small departures from rigidity will have a marked effect on the field. For incompletely rigid bodies like the sun and the earth β will therefore not have exactly the value unity. The results of Hales & Gough (1947) show that the earth's magnetic field decreases on descending a deep mine. This is in conformity with the hypothesis that the magnetic field is a consequence of the gravitational field and cannot be explained on the hypothesis that the magnetic field arises from the presence of electric currents flowing in the Earth's core.

6. DISCUSSION

The theory put forward in the present paper seeks to derive fields directly from the data of experience—namely that physical processes involve 'observables' and 'unobservables'. This requires a five-dimensional framework which is taken as a simple extension of Minkowski's (1908) 'absolute world' by adding a fifth orthogonal direction labelled anti-time. We then make the distinction between observable effects in time (e.g. kinetic energy) and unobservable effects in anti-time (e.g. potential energy). We further adopt the 'common sense' notion of enduring objects in the form of the idealized physical observer O with his measuring system of rigid rulers and clocks and the observed body P idealized as a point mass moving without constraint. The absolute co-ordinate system of the reference framework is defined by means of a hypothetical absolute observer Q .

Three subsidiary conceptions which arise naturally from the fundamental ones are then formulated. The first is that of 'true' straightness defined by straight lines in the Q co-ordinate framework. The unconstrained P -body is associated with a straight time line called the cosmodesic. The second subsidiary conception concerns the rigidity of the measuring system. This is defined by a set of unit vectors at every point. If a constraint (i.e. a field of force) is present these vectors do not coincide with any set of Q -co-ordinates, but diverge from them by small angles. Finally, we have the third notion of 'anti-time' blindness. This is implicit in any five-dimensional theory as de Broglie (1927) pointed out as far back as

1927, but its consequences for field theory have apparently never before been sufficiently explored.

With these simple notions it is possible to set up a field theory which exhibits a remarkable symmetry between gravitational and electrostatic fields. It gives a complete account of electromagnetism and it furnishes a generalized Lagrangian from which the gravitomagnetic effect can be derived and the absence of an associated electrostatic field explained.

These results appear to run counter to the demonstration that the Schwarzschild metric cannot be embedded in a curvature-free fivefold. The explanation is simple: the fundamental equation (3.3) is not that of a metric but of a system of directions at a point. We are not concerned with the metric inherent in the framework, but with peculiarities in the measuring instruments used by O to observe the motion of P . While it is not quite true to say that 'curvature is in the eye of the beholder', our theory shows that a 'truly' straight path may have the appearance of an accelerated motion in a central force field—solely due to the property of rigidity as constraint without deformation and the anti-time blindness of the physical observer.

The merit of a theory largely turns upon the range of observations of which it can give a simple and consistent account. The theory of a five-dimensional world has proved of value outside field theory. It is well known from the works of Kaluza (1921), Klein (1926, 1927) de Broglie (1927), Flint (1942, 1945), Fisher (1929), Rosenfeld (1927), Wilson (1928) and others that there are several attractive features in a five-dimensional wave mechanics—the wave function being associated with a periodicity in the fifth dimension. Throughout Eddington's treatment of fields as a system of relations between pairs of particles, he makes much use of the system of five independent parameters (Eddington 1936 and 1946), but he does not regard the fifth co-ordinate 'the phase co-ordinate' as having the same physical status as the other four.

We hope to show in a future communication that a very satisfactory treatment of wave mechanics can be given within the framework of the present theory. It also suggests the nature of the exchange forces and the meson field theory developed for the purposes of nuclear physics. In our view, general field theory must be regarded as the foundation upon which any cosmology must be constructed in our day. One of the most disappointing features of general relativity has been the failure to build a bridge to join it to the rest of physical science. It has been our aim to construct a scheme so closely related to our common experience that it should be equally applicable to the description of all physical processes. In particular, we feel the need to establish a descriptive scheme which shall enable dynamics, atomic physics and the statistical systems of thermodynamics to be exhibited as a single coherent system. We believe that this is not possible so long as four-dimensional space-time—however generalized—is taken as the framework to which all physical processes must be referred. We have endeavoured to show that a consistent and fruitful world-picture is obtained by extending the space-time framework to a five-dimensional scheme free from the complications of a Riemannian or affine geometry.

APPENDIX. LIST OF SYMBOLS

Q	absolute observer with five orthogonal rectilinear co-ordinates (${}^1Q \dots {}^5Q$). 5Q is called anti-time axis of Q .
${}^1q \dots {}^5q$	mixed co-ordinate set related to ${}^1Q \dots {}^5Q$ by equations (1.1a).
$(X^1, X^2, X^3)(X^4)$	ruler and clock measurements of physical observer O .
$x^1 \dots x^4$	mixed sets of variables related to $X^1 \dots X^4$ by equations (1.1a).
P	freely falling body characterized by a straight track (cosmodesic) in the five-dimensional manifold.
$\frac{1}{2}\pi - \lambda$	angle between Q^5 and the cosmodesic of P .
E	electric charge of P .
m_0	inertial mass of P .
$L_1 \dots L_4$	unit vectors through the origin R along which O 's ruler and clock readings increase at maximum rates respectively.
L_5	unit vector through R along which O 's ruler and clock readings remain stationary (the direction of anti-time for O).
ϖ	unit vector orthogonal to L_1, L_2, L_3, L_4 , having components (${}_1\varpi \dots {}_5\varpi$) with respect to Q .
j, k, n, s, v, ω, p	super- or subscripts which assume the values 1, 2, 3, 4, 5.
μ, ν, σ	super- or subscripts which assume the values 1, 2, 3, 4.
ψ, ξ, η	super- or subscripts which assume the values 1, 2, 3.
	The double suffix summation convention applies to each set over these values.
il_k	purely real or purely imaginary numbers specifying the direction of L_k in the iQ system.
Δx^j	infinitesimal displacements along L_j .
Ω	potential energy of P in a field of force. Ω_g gravitational, Ω_e electromagnetic.
t	time co-ordinate of O .
${}_k l^j$	the co-factor of ${}_k l_j$ in $\ {}^k l_j \ $ divided by the value of this determinant.
r, θ, ϕ	spherical polar co-ordinates in the space of O .
e_ψ	direction cosines in space defined by equation (3.8).
V	velocity of P relative to O , components (V_1, V_2, V_3).
v_1, v_2, v_3	defined as iV_1/c , etc.
ζ	an universal dimensional constant.
U (in § 3)	a velocity defined by the values of the ${}^\nu l_\mu$ in an electromagnetic field (identified with the velocity of the field-producing body with respect to Q and O).
A	electromagnetic four-vector potential defined by (3.16).
(f_1, f_2, f_3)	direction cosines in space defined by (3.17).
ϵ, δ	small quantities.
ϵ_s	the s th differential of ϵ with respect to any ${}^\mu q$.
δ_s	the s th differential of δ with respect to any ${}^\mu q$.
P	the field-force vector in classical field theory.
δM	element of mass of a field-producing heavy body.

- U (in § 5) velocity of δM with respect to Q . (U_1, U_2, U_3) (u_1, u_2, u_3) defined from U in the same way as (V_1, V_2, V_3) (v_1, v_2, v_3) defined from V .
- $L'_\mu x'_\mu$ directions of measurement, and measurements respectively which O would have at R under the field of δM if δM had no velocity with respect to Q .
- ${}^k l'_j$ direction cosines of L'_μ .
- ϖ' direction orthogonal to L'_μ .
- Y^ψ axes along L_ψ . Y^5 axis along ϖ . Y^4 axis orthogonal to Y^ψ and Y^5 .
- Y'^ψ axes along L'_ψ . Y'^5 axis along ϖ' ($=\varpi$). Y'^4 axis orthogonal to Y'^ψ and Y'^5 .
- ${}^\mu \lambda_\nu$ direction cosines of Y^ν with respect to Y'^μ .
- ${}^k \mu_j$ direction cosines of Y'^j with respect to Q^k .
- δ^v_ω the Kronecker delta ($= 0$ when $v \neq \omega$, $= 1$ when $v = \omega$).

REFERENCES

- Blackett, P. M. S. 1947 *Nature*, **159**, 658.
- de Broglie, L. 1927 *J. Phys. Radium*, **8**, 65.
- Cartan, E. 1932 *Actualités Scientifiques et Industrielles* No. 44
- Eddington, A. S. 1924 *Mathematical theory of relativity*, 2nd ed., p. 149. Cambridge University Press.
- Eddington, A. S. 1936 *Relativity Theory of Protons and Electrons*. Cambridge University Press.
- Eddington, A. S. 1946 *Fundamental Theory*. Cambridge University Press.
- Fisher, J. W. 1929 *Proc. Roy. Soc. A*, **123**, 489.
- Flint, H. T. 1942 *Phil. Mag.* **33**, 369.
- Flint, H. T. 1945 *Phil. Mag.* **36**, 365.
- Hales, A. L. & Gough, D. I. 1947 *Nature*, **160**, 746.
- Kaluza, Th. 1921 *S.B. preuss. Akad. Wiss.* p. 968.
- Klein, O. 1926 *Z. Phys.* **37**, 895.
- Klein, O. 1927 *Z. Phys.* **41**, 407.
- Minkowski, H. 1908 *Nachr. Ges. Wiss. Gottingen*, p. 53.
- Mosharrafa, A. M. 1948 *Phil. Mag.* **39**, 728.
- Rosenfeld, L. 1927 *Bull. Acad. Roy. Belge*, p. 304.
- Wilson, H. A. 1923 *Proc. Roy. Soc. A*, **104**, 451.
- Wilson, W. 1928 *Proc. Roy. Soc. A*, **118**, 441.
- Wilson, W. & Cattermole, J. 1939 *Phil. Mag.* **30**, 84.

The heats of formation of free CN and free CH₂, and the relationship between $D(\text{CO})$, $D(\text{CN})$ and $D(\text{N}_2)$

By L. H. LONG, PH.D., *University College, Exeter*

(Communicated by R. G. W. Norrish, F.R.S.—Received 3 November 1948)

The heats of formation of CN and CH₂ have been calculated by a number of independent methods and shown to have the approximate values -92.5 and -70 kcal. respectively. The former value, which is independent of any data on cyanogen, receives close support from a recent estimate of the lattice energies of sodium and potassium cyanides, and corresponds to ~ 114 kcal. for the heat of dissociation of C₂N₂ into two CN radicals.

The information provided by the heat of formation of CH₂ regarding the energies involved in the stepwise dissociation of methane furnishes an indication of the true values for the heat of atomization of graphite and the dissociation energy of carbon monoxide.

The relation linking $D(\text{CO})$, $D(\text{CN})$ and $D(\text{N}_2)$ has been recalculated from the derived value for the heat of formation of CN and utilized as far as possible for deciding on the correct dissociation energies. Of the four values which have been proposed for $D(\text{CO})$ and the three values for $D(\text{N}_2)$, it is possible to eliminate two of the former and one of the latter.

INTRODUCTION

The dimensions of the dissociation energies of CO and N₂ have been conclusively settled in neither case, there being at least four values contested for the former and three for the latter. This being so, one cannot afford to overlook the manner in which they are interrelated, since the fixing of either value should lead, in conjunction with sufficient evidence concerning CN, to the fixing of the other. The relationship takes the form

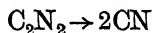
$$D(\text{CO}) = D(\text{CN}) - \frac{1}{2}D(\text{N}_2) + x \text{ kcal.}, \quad (1)$$

where x involves $D(\text{O}_2)$ and the heats of formation of CO and CN from graphite, molecular oxygen and molecular nitrogen respectively.

A relation of this kind has recently been employed by Springall (1947) in an attempt to decide between the various values proposed for the heats of atomization of carbon and nitrogen. Quite independently, Long & Norrish (1946*c*) examined the question of the heat of atomization of carbon, and concluded from several other lines of evidence that the high value around 170 kcal./g.-atom selected by Springall is not a permissible figure. Since the respective authors draw opposing conclusions, some explanation of the discrepancy is called for. In the present paper an explanation is suggested, and further evidence relating to the heats of atomization of carbon and nitrogen discussed.

With respect to the thermochemical values from which x (equation (1)) is derived, the heat of formation of CO (Wagman, Kilpatrick, Taylor, Pitzer & Rossini 1945) is accurately known from the heat of combustion, and the long-accepted spectroscopic value for $D(\text{O}_2)$ (Herzberg 1939) remains unchallenged. The position regarding the heat of formation of the free CN radical is far less satisfactory. Hitherto, this has always been calculated from the heat of formation of C₂N₂ and its heat of dissociation into two CN radicals. Apart from the fact that the heat of formation of C₂N₂ is

probably not known with an accuracy better than ± 10 kcal./mol., no decisive value for the energy absorbed in the process



is available. Published experimental values exhibit wide disagreement, the extreme figures being 77 and 146 kcal.:

	$D(\text{NC-CN})$ (kcal.)
Hogness & Ts'ai (1932)	≈ 127
Kistiakowsky & Gershinowitz (1933)	77 (± 4)
White (1940)	146 (± 4)
Robertson & Pease (1942 <i>a, b</i>)	125-130

Various authors have selected different values on which to base their calculation. Obviously, confirmation for the chosen figure by an independent means is required before confidence can be placed in any one of their widely differing conclusions.

Fortunately, three other lines of experimental evidence are available which have not been applied to this problem before. They are all independent of any work on C₂N₂, and rest upon known facts concerning other molecules containing the CN group. The heats of formation of these molecules are required, and it would be convenient to discuss these first.

THE HEATS OF FORMATION OF CERTAIN CYANIDES

Methyl cyanide

The heat of combustion of CH₃CN per mol. at constant pressure is given as 312.14 kcal. for the gas by Thomsen (1905) and 304.0 kcal. for the liquid by Lemoult (1909). These figures have been corrected by Kharasch (1929) for more accurate physical constants, as follows:

	heat of combustion (kcal.)
CH ₃ CN gas (Thomsen)	310.4 (18° C)
CH ₃ CN liquid (Lemoult)	302.4 (room temp.)

Since the heat of vaporization at the relevant temperature is 8.05 kcal. (Heim 1933), agreement is excellent, in fact, fortuitously good. This value for the heat of combustion leads to -19.8 kcal. for the heat of formation of gaseous CH₃CN at 18° C, which figure corresponds to -21.4 kcal. at 0° K.

Iodine cyanide

The only data for ICN are due to Berthelot (1875), who measured the heat of formation from potassium cyanide and elementary iodine in aqueous solution, and also the heat of solution of ICN. Berthelot's figures have been combined with modern thermochemical data at the National Bureau of Standards, Washington, to give -40.4 kcal. for the molecular heat of formation of solid ICN from its elements in their standard states at 25° C.* This figure, together with the heats of sublimation

* F. D. Rossini, private communication.

of iodine (Gillespie & Fraser 1936) and ICN (Ketelaar & Kruyer 1943), yields -47.3 kcal. at 25°C corresponding to -47.1 kcal. at 0°K for the heat of formation of gaseous ICN from *gaseous* I_2 , N_2 and graphite, this being the quantity which will be required subsequently. (In the absence of experimental data for the heat capacity of solid ICN above 25°C , the fact that measurements of Ketelaar & Kruyer refer to the temperature range 64 to 153°C and not to 25°C has been ignored. The error thereby introduced would not amount to more than a few tenths of a kilocalorie.)

Hydrogen cyanide

The heat of formation of gaseous HCN can be calculated from the heat of combustion, for which Berthelot (1881) and Thomsen (1905) give figures differing by only a few tenths of 1 %. According to Kharasch (1929), Thomsen's figures are to be preferred for gases and very volatile compounds, those of Berthelot usually being too high. Kharasch further recommends applying a correction of -0.4% to Thomsen's figures to bring them into accord with modern determinations. This reduces Thomsen's figure from 158.62 to 158.0 kcal., which with modern data gives -29.8 kcal. for the heat of formation of gaseous HCN at 18°C , or -29.9 kcal. at 0°K .

Cyanogen

From published and rather discordant data on the heat of combustion of C_2N_2 , Bichowsky & Rossini (1936) calculate the molecular heat of formation of gaseous cyanogen from diamond and hydrogen to be -71 kcal. In this the figure 94.45 kcal. was utilized for the heat of formation of CO_2 from diamond. Employing instead the figure 94.05 kcal. for the heat of formation of CO_2 from graphite as reference state, the value for C_2N_2 becomes -71.8 kcal. at 18°C or -71.4 kcal. at 0°K .

The reduction of the foregoing heats of formation to absolute zero was accomplished with the aid of thermodynamical data from Wagman *et al.* (1945) for graphite, hydrogen and nitrogen, Giaque (1931) for gaseous iodine, Thompson (1941) for methyl cyanide and cyanogen, Stevenson (1939) for iodine cyanide and Gordon (1937) for hydrogen cyanide.

THE HEAT OF FORMATION OF THE FREE CN RADICAL

The long wave-length limits of the second regions of continuous absorption of CH_3CN and ICN are 1600 \AA (Herzberg & Scheibe 1930; Cutler 1948) and 2100 \AA (Badger & Woo 1931; Mooney & Reid 1931, 1932) respectively, these regions corresponding to photodissociation in which, as demonstrated by fluorescence, CN radicals are produced in the excited $B\ ^2\Sigma$ state (Terenin & Neuimin 1934, 1935; Neuimin & Terenin 1936; Yakovleva 1938). The respective limiting energies are 178.6 and 136.1 kcal. Subtracting in each case the excitation energy of the CN radical, namely, 73.6 kcal., the values obtained for $D(\text{CH}_3\text{—CN})$ and $D(\text{I—CN})$ are 105.0 and 62.5 kcal. respectively. If the products of photodissociation carry away excess energy with them, whether translational, rotational or vibrational, then these figures must be reduced correspondingly. From either value the heat of formation of the CN radical may be deduced.

For the energy required to remove the first hydrogen atom from the methane molecule at 0° K, Kistiakowsky & Van Artsdalen (1944) give 100.8 ± 1 kcal. This quantity involves a figure for $D(\text{HBr})$ which needs some correction. From the molecular heat of formation of HBr and the heat of vaporization of bromine at 25° C, 8.66 and 7.34 kcal. respectively (Rossini, Wagman, Evans, Blau & Levine 1947), the heat of formation of HBr from hydrogen and gaseous bromine is 12.33 kcal., which, with thermodynamical data for H₂ (Wagman *et al.* 1945) and for Br₂ and HBr (Gordon & Barnes 1933), corresponds to 12.22 kcal. at 0° K. Combining this with $D(\text{H}_2)$ and $D(\text{Br}_2)$ (Herzberg 1939), $D(\text{HBr})$ becomes 86.5 kcal., which corrects the figure given by Gaydon (1948). The figures employed by Kistiakowsky & Van Artsdalen is 85.8 kcal., so that $D(\text{CH}_3\text{—H})$ becomes 101.5 kcal. at 0° K. This can be combined with the heat of formation of methane at 0° K (Prosen, Pitzer & Rossini 1945) as follows:

	kcal.
$2\text{C}(\text{graphite}) + 1\frac{1}{2}\text{H}_2 + \frac{1}{2}\text{N}_2 \rightarrow \text{CH}_3\text{CN}(\text{gas})$	— 21.4
$\text{CH}_3\text{CN}(\text{gas}) \rightarrow \text{CH}_3 + \text{CN}(X^2\Sigma)$	— 105.0
$\text{CH}_3 + \text{H} \rightarrow \text{CH}_4(\text{gas})$	+ 101.5
$\frac{1}{2}\text{H}_2 \rightarrow \text{H}$	— 51.6
$\text{CH}_4(\text{gas}) \rightarrow \text{C}(\text{graphite}) + 2\text{H}_2$	— 15.987
$\text{C}(\text{graphite}) + \frac{1}{2}\text{N}_2 \rightarrow \text{CN}(X^2\Sigma)$	— 92.5

The heat of formation of the CN radical can also be calculated from the heat of formation of ICN, $D(\text{I—CN})$ and $D(\text{I}_2)$ (Herzberg 1939):

	kcal.
$\text{C}(\text{graphite}) + \frac{1}{2}\text{N}_2 + \frac{1}{2}\text{I}_2(\text{gas}) \rightarrow \text{ICN}(\text{gas})$	— 47.1
$\text{ICN}(\text{gas}) \rightarrow \text{I}(^2P_{3/2}) + \text{CN}(X^2\Sigma)$	— 62.5
$\text{I}(^2P_{3/2}) \rightarrow \frac{1}{2}\text{I}_2(\text{gas})$	+ 17.8
$\text{C}(\text{graphite}) + \frac{1}{2}\text{N}_2 \rightarrow \text{CN}(X^2\Sigma)$	— 91.8

These two calculations from the absorption spectra confirm one another to well within the limits of experimental error. Presumably the spectra of ClCN and BrCN would exhibit similar regions of continuous absorption, but these have not been examined at sufficiently short wave-lengths.

The third method of calculating the heat of formation of CN depends upon an extrapolation of vibrational levels in the absorption spectrum of HCN (Funke & Lindholm 1937) according to the method of Rydberg. The extrapolation leads to 41,000 cm.⁻¹ or 117 kcal. for $D(\text{H—CN})$. An indication of the reliability of the method is provided by the very similar case of acetylene (Funke & Lindholm 1937), where the value calculated for $D(\text{H—C}_2\text{H})$ is virtually identical with that obtained from the continuum onset at 2350 Å (Cherton 1942, 1943). (Here the result of the Rydberg extrapolation is some 25 % below that provided by the Birge-Sponer method which, as generally recognized, normally gives a figure considerably in excess of the true value.) Again it is possible to deduce the heat of formation of CN:

	kcal.
$\text{C}(\text{graphite}) + \frac{1}{2}\text{H}_2 + \frac{1}{2}\text{N}_2 \rightarrow \text{HCN}(\text{gas})$	— 29.9
$\text{H}(^2S) \rightarrow \frac{1}{2}\text{H}_{2g}$	+ 51.6
$\text{HCN}(\text{gas}) \rightarrow \text{H}(^2S) + \text{CN}(X^2\Sigma)$	— 117.2
$\text{C}(\text{graphite}) + \frac{1}{2}\text{N}_2 \rightarrow \text{CN}(X^2\Sigma)$	— 95.5

The three independent figures thus obtained for the molecular heat of formation of free CN are in satisfactory agreement in view of the rather large possible errors:

	kcal.
continuum onset for CH_3CN	-92.5
continuum onset for ICN	-91.8
Rydberg extrapolation for HCN	-95.5

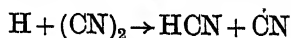
Of these, the first two would normally be more rightly regarded as lower limits (algebraically speaking), but this would be absolutely certain only if the CN fluorescence had been observed right up to the long wave-length limits of the respective continua of CH_3CN and ICN. Until this point has been examined in detail, there is also a possibility that these figures err somewhat on the high side. In the case of iodine cyanide, however, Yakovleva (1938) has provided further evidence which limits this possibility. When ICN was illuminated with the light from a zinc spark containing the wave-lengths 2100, 2064 and 2025 Å no fluorescence was observed. With an aluminium spark containing the lines 1990, 1935, 1864 and 1854 Å, however, fluorescence was obtained. Here the type of emission was such as to indicate that the sum of the vibrational and rotational energies of the CN radicals produced was small, in contrast to the internal energies of the radicals produced by the hydrogen lamp. Allowing for a certain amount of translational energy, this is strong evidence that the figure derived from the continuum limit around 2100 Å will not be very many kilocalories above or below the true value.

In the following calculations the value -92.5 kcal. will be adopted for the heat of formation of CN, less because it is the intermediate of the three figures than because the heat of formation of methyl cyanide is known with rather greater precision than is the case for the other two compounds. From this value and the heat of formation of cyanogen, $D(\text{NC-CN})$ at 0° K may be estimated:

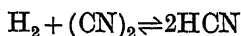
	kcal.
$2\text{C}(\text{graphite}) + \text{N}_2 \rightarrow 2\text{CN}(X^2\Sigma)$	-185.0
$\text{C}_2\text{N}_2(\text{gas}) \rightarrow 2\text{C}(\text{graphite}) + \text{N}_2$	+ 71.4
$\text{C}_2\text{N}_2(\text{gas}) \rightarrow 2\text{CN}(X^2\Sigma)$	-113.6

From this, $D(\text{NC-CN})$ is seen to assume the value 113.6 kcal., a very reasonable figure lying approximately midway between the highly discordant extremes in the literature. No high accuracy is claimed for this figure. Nevertheless, the foregoing evidence is scarcely compatible with the extreme values 77 (Kistiakowsky & Gershinowitz 1933) and 146 kcal. (White 1940). In going to press, it is interesting to note that Glockler (1948*a*) also rejects these two values from a study of the respective lattice energies of sodium cyanide and potassium cyanide in conjunction with the Born-Haber cycle. The values Glockler thus obtains for $D(\text{NC-CN})$ are 117.6 and 119.6 kcal. respectively, in good agreement with the figure calculated here, and hence with the value derived for the heat of formation of CN.

The reasonable dimension of the figure obtained for $D(\text{NC-CN})$ is also indicated by a study of the reaction between cyanogen and hydrogen at high temperatures (Robertson & Pease 1942*b*), since, as pointed out by Steacie (1946), the process



must have a low activation energy in order that it may serve as a chain-carrying step. This could not be the case if $D(\text{NC—CN})$ were much greater than $D(\text{H—CN})$ (c. 117 kcal.). In harmony with this, Robertson & Pease observed that the equilibrium



lies experimentally far to the right, indicating that $D(\text{H—CN})$ is appreciably greater than the mean of $D(\text{H}_2)$ and $D(\text{NC—CN})$. This would put an upper limit of about 130 kcal. on $D(\text{NC—CN})$. These points provide further evidence against White's high value which Springall (1947) favours. The figure derived in the present paper is further seen to be reasonable when compared with $D(\text{CH}_3\text{—CN})$ and $D(\text{CH}_3\text{—CH}_3)$, which latter value may be calculated from the heats of formation of methane and ethane at 0° K (Prosen, Pitzer & Rossini 1945), $D(\text{H}_2)$ and the figure 101.5 kcal. for $D(\text{CH}_3\text{—H})$ (see previously):

	kcal.
$D(\text{CH}_3\text{—CH}_3)$	84.3
$D(\text{CH}_3\text{—CN})$	105.0
$D(\text{NC—CN})$	113.6

The middle member has a rather higher dissociation energy than the mean value of the other two. This is usually observed for similar series. Here it is conditioned by the heats of formation of the three compounds concerned. Other workers have argued, in support of a higher value for $D(\text{NC—CN})$, that conjugation effects within the molecule will greatly strengthen the C—C bond. This is doubtless an important factor, but Walsh (1948) has drawn attention to another factor acting in the opposite direction. In the rather similar molecules glyoxal and diacetyl, even though conjugation of the two carbonyl groups results in the shortening of the central C—C bond, polarity effects offset a corresponding increase in the bond energy and dissociation energy of the C—C link, which, in the case of diacetyl at least, is more easily ruptured than the C—C link in ethane. Similarly in cyanogen, as predicted by Walsh, the polarity of the $\text{C}\equiv\text{N}$ groups is such as to withdraw electrons from the C—C bond, so reducing the overlapping of bonding electrons and hence the bond energy. A confirmatory indication is supplied by the stretching force constants of the C—C bonds. Whereas those for C_2H_6 and CH_3CN , as calculated by Linnett (1940, 1941), are 4.53 and 5.3×10^5 dyne/cm. respectively, the corresponding value for C_2N_2 , namely, 5.22×10^5 dyne/cm. (Herzberg 1945), differs little from that for CH_3CN . We are here dealing with dissociation energies and not bond energies, but even allowing for the difference in reorganization energies of the CH_3 and CN radicals, there is again no indication that $D(\text{NC—CN})$ will be much greater than $D(\text{CH}_3\text{—CN})$.

These conclusions constitute one point of difference with Springall (1947) who selects the high value 142 kcal. for $D(\text{NC—CN})$. Goldfinger (1947) also criticizes Springall on this point. The other major point of difference is the value adopted for $D(\text{CN})$, which will be discussed later.

THE HEAT OF FORMATION OF METHYLENE

It would be relevant at this point to bring forward certain evidence regarding the heat of formation of methylene, CH_2 , since this provides an indication of the true

value of $D(\text{CO})$. The latter quantity is related to the heat of sublimation of carbon (graphite) into ground-state atoms, L_1 , by the equation

$$D(\text{CO}) = L_1 + 85.78 \text{ kcal.} \quad (2)$$

Now, from equations given previously (Long & Norrish 1946c),

$$L_1 = a + b + c + d - 226.1 \text{ kcal.}$$

at 25° C, where a , b , c and d are the respective energies required for the processes:



At 0° K the equation becomes

$$L_1 = a + b + c + d - 222.4 \text{ kcal.} \quad (3)$$

Owing to lack of information concerning the value of b , this method has not so far been successfully applied to the evaluation of L_1 .

One method of evaluating the heat of formation of CH_2 and hence b is provided by the investigation of the *homogeneous* decomposition of methane. Failure to distinguish between the homogeneous and heterogeneous processes has led to controversy in the past. Many investigators have studied the heterogeneous decomposition, but Kassel (1932) seems to be the only one to have studied the homogeneous process. He carried out his experiments in a quartz vessel at pressures between 1 and 40 cm. in the temperature range 700 to 850° C. The reaction velocity was not appreciably affected by increasing the surface area twentyfold. The decomposition was found to be unimolecular with an activation energy of 79 ± 6 kcal., the final products being carbon, hydrogen and a trace of oily matter. There was, however, an unexplained induction period for some of the experiments. Of the two possible initial processes,



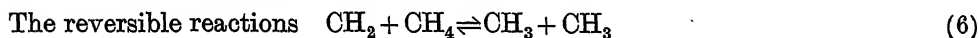
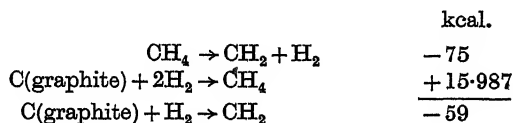
and



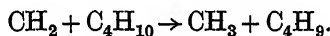
the former seems to be excluded on energetic grounds, since for a homogeneous process the activation energy cannot be less than the energy absorbed, which for process (4) is 101.5 kcal. at 0° K (see the previous section). The initial act must consequently be process (5), conclusions which are supported by Barrow, Pearson & Purcell (1939). On the other hand, there is a fair amount of evidence to show that the heterogeneous decomposition of methane is mainly concerned with process (4). Thus Eltenton (1947) in mass-spectrograph experiments with methane observed CH_3 and not CH_2 ; but with a deactivated (as distinct from an activated) surface and a pressure below 0.3 mm. he also failed to observe CH_3 , even at 1100° C, an observation which may in some way be connected with the induction period observed by Kassel. Eltenton thus connects process (4) with the heterogeneous decomposition.*

* G. C. Eltenton, private communication.

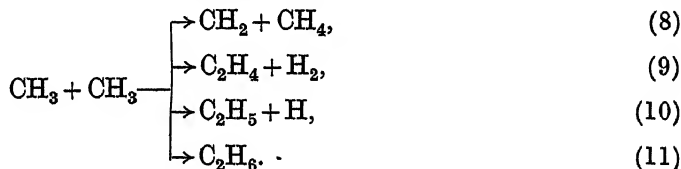
Kassel's activation energy for process (5), the one with which we are here concerned, corresponds to 75 kcal. at 0° K, and -59 kcal. for the heat of formation of CH₂:



have already been used by Wicke (1942, 1945) for deriving the heat of formation of methylene. Experimental evidence for reaction (6) has been obtained by Rosenblum (1941). Pearson, Purcell & Saigh (1938) also conclude that it occurs at high temperatures, though not appreciably at room temperature since methane scarcely affects the life of CH₂. These facts are in harmony with its somewhat endothermic nature, and imply an activation energy in the range 10 to 20 kcal. The experiments of Rice & Glasebrook (1934) concerning the decomposition of diazomethane in the presence of butane also lead to an activation energy of 15 ± 5 kcal. for the very similar reaction



The reaction between two methyl radicals may follow the paths indicated below:



Process (10) is endothermic and less likely than (8) or (9). This is born out by the experiments of Paneth, Hofeditz & Wunsch (1935) who examined the products obtained from methyl radicals, both with helium as a carrier gas and in the absence of a carrier gas. Process (10) can be eliminated since it would have resulted in the formation of propane and butane, which was not observed. Further, since the C₂H₄:CH₄ ratio was never too great to be explained entirely by process (8) (the reverse of reaction (6)) followed by the recombination of CH₂ to ethylene on the walls, the indication is that reaction (9) is unimportant. Thus the main gas-phase reaction appears to be the disproportionation process (8), with process (11) occurring on the walls. Reactions between two radicals, unless they are endothermic, are usually found to have low activation energies. However, that of reaction (8) is not negligible since the decomposition of silver methyl (Semerano & Riccoboni 1941) produces ethane only and no methane or ethylene at low temperatures, in contrast to higher temperatures. Furthermore, Paneth *et al.* (1935) succeeded in increasing the half-life of CH₃ to about 0.1 sec. In accordance with the foregoing, 8 ± 4 kcal. would appear to be a fair estimate of the activation energy of reaction (8).

Neither reaction (7) nor the back process requires a high activation energy. Rosenblum (1938, 1941) finds that methylene is much less stable towards hydrogen than towards methane. From this it follows that process (7) has a lower activation

energy than process (6), so that 10 ± 5 kcal. may be regarded as an approximate estimate for the former. In addition, the temperature dependence observed by Rosenblum for the formation of saturated hydrocarbons requires a low activation energy for reaction (7). For the back reaction, which involves an atom and a free radical, a very low activation energy is to be expected. Accordingly, Trenner, Morikawa & Taylor (1937) postulate the process $\text{CH}_3 + \text{D} \rightarrow \text{CH}_2 + \text{HD}$ as a possible rapid step in the deuterization of methane, the alternative process $\text{CH}_3 + \text{D} \rightarrow \text{CH}_2\text{D} + \text{H}$ being regarded as unlikely by Gorin, Kauzmann, Walter & Eyring (1939). Steacie (1946) accepts 5 kcal. as an upper limit for the activation energy of the reverse of reaction (7), so that 2.5 ± 2.5 kcal. may here be regarded as a safe figure.

The differences between the respective activation energies of the forward and back processes for the two reactions (6) and (7) lead to -7 and -7.5 kcal. respectively for the approximate heats of reaction. These conclusions are similar to those of Wicke (1945), and also receive support from Barrow *et al.* (1939) who consider both reactions are nearly thermoneutral. These figures correspond to -77 and -78 kcal. respectively for the heat of formation of CH_2 :

	kcal.
$\text{C}(\text{graphite}) + 2\text{H}_2 \rightarrow \text{CH}_4$	+ 15.987
$2\text{CH}_3 \rightarrow \text{CH}_4 + \text{CH}_2$	+ 7
$2\text{CH}_4 \rightarrow 2\text{CH}_3 + 2\text{H}$	- 203.0
$2\text{H} \rightarrow \text{H}_2$	+ 103.2
$\text{C}(\text{graphite}) + \text{H}_2 \rightarrow \text{CH}_2$	- 77
<hr/>	
$\text{C}(\text{graphite}) + 2\text{H}_2 \rightarrow \text{CH}_4$	+ 15.987
$\text{CH}_3 + \text{H} \rightarrow \text{CH}_2 + \text{H}_2$	+ 7.5
$\text{CH}_4 \rightarrow \text{CH}_3 + \text{H}$	- 101.5
$\text{C}(\text{graphite}) + \text{H}_2 \rightarrow \text{CH}_2$	- 78

Finally, there is evidence from the ultra-violet absorption spectrum of ketene. This exhibits a continuum over the range 2600 to 3700 Å in which diffuse bands are to be distinguished, this region corresponding to the photolytic decomposition process $\text{CH}_2\text{CO} \rightarrow \text{CH}_2 + \text{CO}$ (Norrish, Crone & Saltmarsh 1933; Norrish 1934). These results have received independent confirmation (Ross & Kistiakowsky 1934). The quantum yield is nearly unity, but falls off near the long wave-length limit. The point of onset of the continuum puts an upper limit of 77 kcal. on the energy required to effect the photo-decomposition. The molecular heat of formation of gaseous ketene may be calculated from the heat of reaction with dilute caustic soda (Rice & Greenberg 1934) and thermochemical data for sodium hydroxide and sodium acetate in solution (Bichowsky & Rossini 1936) to be 15 kcal. at 18° C, corresponding to 14 kcal. at 0° K. Combining this figure with the heat of formation of carbon monoxide (Wagman *et al.* 1945), the figure -91 kcal. is obtained as a lower limit (algebraically speaking) for the heat of formation of CH_2 :

	kcal.
$2\text{C}(\text{graphite}) + \text{H}_2 + \frac{1}{2}\text{O}_2 \rightarrow \text{CH}_2\text{CO}(\text{gas})$	+ 14
$\text{CH}_2\text{CO}(\text{gas}) \rightarrow \text{CH}_2 + \text{CO}$	- 77
$\text{CO} \rightarrow \text{C}(\text{graphite}) + \frac{1}{2}\text{O}_2$	- 27.202
$\text{C}(\text{graphite}) + \text{H}_2 \rightarrow \text{CH}_2$	- 91

Owing to the rather large inaccuracies involved, close agreement is not to be expected between the results from the various methods of calculating the heat of formation of CH₂. All the evidence is compatible with the figure -70 ± 15 kcal., and there is no indication of a decidedly different value. From this figure, *b* (equation (3)) is found to be 88 kcal.:

	kcal.
CH ₄ → C(graphite) + 2H ₂	- 15·987
CH ₃ + H → CH ₄	+ 101·5
H ₂ → 2H	- 103·2
C(graphite) + H ₂ → CH ₂	- 70
CH ₃ → CH ₂ + H	- 88

The value of *d*, that is, *D*(CH), has been calculated by Herzberg (1939) from observations of band spectra by Shidei (1936) to be 80·0 kcal., a value in excellent accord with the expectations of Heimer (1932), and also with the general trend of C—H bond energies as related to bond lengths and force constants (Walsh 1947*a*, 1948). The value for *c*, as pointed out formerly (Long & Norrish 1946*c*), could as a first approximation be put equal to *d*. From theoretical considerations, Voge (1936) calculates that *c* is 0·05 eV in excess of *d*. Accordingly, 81 kcal. may be taken as the 'most probable' value for *c*, a figure not likely to be in error by many kilocalories. These values, in conjunction with equations (2) and (3), lead to 128 kcal. for *L*₁ and 214 kcal. for *D*(CO). The actual values given by the individual methods are as follows:

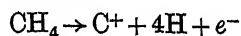
	<i>L</i> ₁ (kcal.)	<i>D</i> (CO) (kcal.)
(i) homogeneous decomposition of CH ₄	117	203
(ii) heat of reaction of CH ₂ + CH ₄ ⇌ 2CH ₃	135	221
(iii) heat of reaction of CH ₂ + H ₂ ⇌ CH ₃ + H	136	222
(iv) photo-decomposition of ketene	≥ 149	≥ 235

Methods (i), (ii) and (iii), individually as well as collectively, seem to be entirely incompatible with the high value 256·1 and the low value 159·7 kcal. which have been proposed for *D*(CO) (see table 1, p. 74).

It needs at this point to be mentioned that a recent investigation by an effusion method (Brewer, Gilles & Jenkins 1948) appears to indicate a relatively low vapour pressure for graphite and hence support the high value for *D*(CO), which is in definite conflict with the foregoing evidence. However, it has been pointed out elsewhere (Long 1948) that the results of Brewer *et al.* are also in conflict with other available evidence concerning the direct measurement of the *equilibrium* vapour pressure and triple point of graphite, and that these investigators themselves make observations concerning the rapid volatilization of graphite which are scarcely compatible with their own temperature-pressure relation and its implied low vapour pressure. Because of the importance of all direct evidence, further comments are called for. Even when the first of the seven independent measurements which Brewer *et al.* record is ignored as discordant, the remaining six results are seen to indicate vapour pressures which, allowing for the slight temperature modulation, vary among themselves by a factor of approximately 10. Yet, from the results of

varying the area of the effusion hole by a factor of only 2.77, it is concluded that the accommodation coefficient for graphite is sufficiently large to ensure the attainment of true equilibrium vapour pressures inside a crucible with an effusion hole of diameter equal to one-quarter of the internal diameter of the crucible. The absence of a more thorough investigation of this fundamental point, upon which the validity of the results entirely depends, is obviously a serious omission, the more so because unpublished work of Johnston & Marshall referred to by Herzberg (1942) indicates an 'exceedingly small' accommodation coefficient. Brewer *et al.* further tacitly assume that *all* of the carbon vapour impinging on a cold platinum surface adheres to it, and that a true Maxwellian distribution for the vapour effusing (as demanded by the effusion equation employed) is realized under conditions where the mean free path is long compared with the dimensions of the enclosure. For a number of reasons, therefore, it would be unwise to place great emphasis on these results before they are confirmed by independent measurements of the equilibrium vapour pressure of graphite, preferably over a wider temperature range. (These criticisms do not apply to the determination of $D(C_2)$ also described by Brewer *et al.*)

Apart from this instance, the indication of the evidence discussed in this section, namely that neither of the extreme values proposed for $D(CO)$ is permissible, receives support from the bulk of other direct evidence concerning the equilibrium vapour pressure of graphite (Long 1948; Long & Norrish 1946 *b, c*). The forementioned photolysis of ketene (method (iv)) is also particularly valuable evidence against the high value, and may be compared with the observation of Faltings, Groth & Harteck (1938) that ultra-violet radiation of wave-length 1295 Å effects the photo-decomposition of carbon monoxide with a quantum yield of unity, thus placing an upper limit of 221 kcal. on $D(CO)$, corresponding to 135 kcal. on L_1 . In this respect, Schmid & Gerö (1946) have brought forward three independent arguments against an attempt (Gaydon & Penney 1942) to reconcile the latter evidence with the high value (256.1 kcal.) for $D(CO)$. Chemically speaking, a value as high as this for the energy required to rupture a double bond (Long & Walsh 1947) would in any case be without precedent. Again, the process



is observed in electron-impact experiments at 26.7 ± 0.7 eV (Smith 1937), which places an upper limit on the energy of atomization of methane and hence on L_1 . From this and three other processes of like type referred to by Hagstrum (1947), upper limits for L_1 have been calculated, all of which lie in the range 131 to 141 kcal. None of the evidence discussed in this section gives a sufficiently accurate figure to be able to decide conclusively between the alternatives 210.75 and 221 kcal. which have also been suggested for $D(CO)$ (see table 1), although it is rather difficult to reconcile that from method (i) with the latter alternative. It is emphasized that the various calculations of L_1 and $D(CO)$ by methods (i), (ii), (iii) and (iv) all depend upon a theoretically derived value for the step $CH_2 \rightarrow CH + H$, but are otherwise entirely experimental. An error of several kilocalories here would not alter the general conclusions, which are in complete accord with the evidence cited previously (Long & Norrish 1946 *a, b, c*); nor likewise would the eventuality, discussed as a possibility

by Walsh (1947*b*), that the triplet state of CH₂ should turn out to be about 5 kcal. lower in energy than the singlet state (Voge's calculation of the quantity *c* referring to the latter).

THE RELATIONSHIP BETWEEN $D(\text{CO})$, $D(\text{CN})$ AND $D(\text{N}_2)$

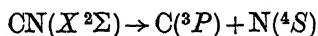
From the derived heat of formation of CN, the value of x in equation (1) may be evaluated at 178.3 kcal. as follows:

	kcal.	
$\text{CO}(X^1\Sigma) \rightarrow \text{C}(^3P) + \text{O}(^3P)$	– $D(\text{CO})$	
$\text{C}(^3P) + \text{N}(^4S) \rightarrow \text{CN}(X^2\Sigma)$	+ $D(\text{CN})$	
$\frac{1}{2}\text{N}_2 \rightarrow \text{N}(^4S)$	– $\frac{1}{2}D(\text{N}_2)$	
$\text{CN}(X^2\Sigma) \rightarrow \text{C}(\text{graphite}) + \frac{1}{2}\text{N}_2$	+ 92.5	
$\text{O}(^3P) \rightarrow \frac{1}{2}\text{O}_2$	+ 58.6	
$\text{C}(\text{graphite}) + \frac{1}{2}\text{O}_2 \rightarrow \text{CO}(X^1\Sigma)$	+ 27.202	
$D(\text{CO}) = D(\text{CN}) - \frac{1}{2}D(\text{N}_2)$	+ 178.3	(1 <i>a</i>)

Equation (1*a*) is entirely experimental.

The spectrum of the CN radical has been examined by several investigators, but most recently and in greatest detail by Schmid, Gerö & Zemplén (1938). By following a series of perturbations, the $A^2\Pi$ state has been observed up to the 30th vibrational sublevel. A short extrapolation leads to a convergence limit at 60,500 cm.^{–1} or 172.8 kcal. above the ground state of CN. The convergence of the upper $B^2\Sigma$ state is extrapolated to 65,500 cm.^{–1}, that is, 5000 cm.^{–1} higher, the stated accuracy in each case being ± 1000 cm.^{–1}.

Since neither carbon nor nitrogen has an excited level in the neighbourhood of 5000 cm.^{–1} above the respective ground state, the nearest being C: 1D at 10,192 cm.^{–1}, it would follow that the $A^2\Pi$ state of CN is not derived from ground-state atoms. Accordingly, Schmid *et al.* (1938) conclude that the relevant combination of atomic states is C: 3P + N: 2P . If this interpretation is correct, the energy required for the process



is 31,660 cm.^{–1} or 90.5 kcal., in contradistinction to the much higher value adopted by Springall (1947). (Employing the newly observed level for the 5S state of carbon (Shenstone 1947), the value obtained from the $B^2\Sigma$ convergence to C: 5S + N: 4S at 65,500 cm.^{–1} differs from this by only ~ 100 cm.^{–1}.) Substituting this value for $D(\text{CN})$ in equation (1*a*), the latter becomes:

$$D(\text{CO}) = 268.8 \text{ kcal.} - \frac{1}{2}D(\text{N}_2). \quad (1b)$$

The values of $D(\text{CO})$ and $D(\text{N}_2)$ which are at present receiving support are given in table 1. For the sake of consistency, the spectroscopic values here quoted have been recalculated from the predissociation limits at $89,620 \pm 50$ cm.^{–1} for CO (Schmid & Gerö 1935; Gerö 1936) and $97,960 \pm 40$ cm.^{–1} for N₂ (Büttenbender & Herzberg 1934) by subtracting the relevant excitation energies and converting to kilocalories. The conversion factors employed are those compiled by Herzberg (1944) from the review of Birge (1941). The figures quoted for Valatin are based on the work of Schmid & Gerö (1937, 1943). In the case of CO, these authors consider that the products of

dissociation for the limit at $89,620\text{ cm.}^{-1}$ are either $\text{C:}^5\text{S} + \text{O:}^3\text{P}$ or $\text{C:}^3\text{P} + \text{O:}^1\text{S}$. The difference amounts to only a few cm.^{-1} , the figure quoted here being based on the former assumption. Hagstrum's figures are the electron-impact values, that for nitrogen being essentially the same as Herzberg's spectroscopic value.

TABLE 1

	$D(\text{CO})$ (kcal.)	$D(\text{N}_2)$ (kcal.)
Gaydon (1947)	256.1	225.0
Hagstrum (1947)	~ 221	~ 171
Herzberg (1939)	210.75	170.1
Valatin (1946a)	159.7	115.1

Of the possible values listed in table 1, only two pairs will provide a reasonable satisfaction to equation (1b), namely,

$$D(\text{CO}) = 210.75 \quad \text{and} \quad D(\text{N}_2) = 115.1;$$

$$D(\text{CO}) = 159.7 \quad \text{and} \quad D(\text{N}_2) = 225.0.$$

(In neither case does the discrepancy with the equation amount to as much as 4 kcal.)

This result is surprising and brings certain difficulties with it. Whereas, from the evidence already discussed, 210.75 kcal. is a permissible value for $D(\text{CO})$, the low value 115.1 kcal. for $D(\text{N}_2)$ is difficult to accept on the grounds that it leads to impossibly low dimensions for the bond energies in many nitrogen compounds, particularly that for the N—N bond energy in hydrazine. On the other hand, the alternative pair of values entails the very low figure 73.9 kcal./g.-atom for the heat of atomization of graphite, a figure at serious variance with the evidence already brought forward, as well as with direct measurement of the vapour pressure of carbon, which indicates a much higher value.

This forces us to look once again at the means we have employed in arriving at these alternative conclusions. In calculating equation (1b) three experimental values have been adopted, namely those for $D(\text{O}_2)$, $D(\text{CN})$ and the heat of formation of the CN radical. Since the last-named value has been derived by three independent methods which agree to within the limits of experimental error, both among themselves and with the calculations of Glockler (1948a), and since the value of $D(\text{O}_2)$ appears to be satisfactorily established, of the three quantities adopted that for $D(\text{CN})$ is the most likely to be in error. Although the scientific world has not yet been presented with an alternative analysis of the CN spectrum comparable in detail with that given a decade ago by Schmid *et al.* (1938), it would be useful at this point to reverse the calculations we have made and deduce from equation (1a) the dimension of $D(\text{CN})$ corresponding to each of the possible combinations of $D(\text{CO})$ and $D(\text{N}_2)$. The results are given in table 2. The possible values of $D(\text{CN})$ obtained from the lowest known convergence limit at $60,500\text{ cm.}^{-1}$ by assuming different products of dissociation are given in table 3.

TABLE 2. VALUES OF $D(\text{CN})$ CALCULATED FROM POSSIBLE COMBINATIONS OF $D(\text{CO})$ AND $D(\text{N}_2)$ (IN KCAL.)

$D(\text{CO}) \backslash D(\text{N}_2)$	225.0	170.1	115.1
256.1	(a) 190.3	(b) 162.8	(c) 135.3
221	(d) 155	(e) 128	(f) 100
210.75	(g) 144.9	(h) 117.5	(i) 90.0
159.7	(j) 93.9	(k) 66.4	(l) 38.9

TABLE 3. POSSIBLE VALUES OF $D(\text{CN})$ FROM THE CONVERGENCE LIMIT AT $60,500 \pm 1000 \text{ cm.}^{-1}$

dissociation products		$D(\text{CN})$ (kcal.)
carbon	nitrogen	
3P	4S	172.8
1D	4S	143.8
3P	2D	117.9
1S	4S	111.0
3P	2P	90.5

The evidence discussed here in the section on the heat of formation of methylene points unequivocally against cases *a*, *b*, *c*, *j*, *k* and *l* (table 2). A comparison of tables 2 and 3 shows that cases *a*, *k* and *l* are doubly ruled out, the former because the highest possible value of $D(\text{CN})$ is too low, and the latter two for precisely the reverse reason, there being no likelihood that CN dissociates into products which are more highly excited than those suggested by Schmid *et al.* (1938). It would thus seem, quite independently of the correct dimension of $D(\text{CN})$, that the values for $D(\text{CO})$ and $D(\text{N}_2)$ proposed by Gaydon are not mutually compatible, nor likewise those favoured by Valatin. Among the main grounds of support for case *a* have been theoretical ones based on the Non-crossing Rule (Gaydon & Penney 1945). But a rigid application of this rule involves an identification of the potential curves of the unperturbed electronic states of the molecules concerned with those of the eigenvalues provided by the classical two-centre model. As soon as the relative motions of the nuclei are considered, this identification is no longer permissible (Valatin 1946*b*). It is therefore not surprising that the Non-crossing Rule has led to incorrect conclusions, especially as other diatomic molecules are known which do not conform to it.

Of the remaining six cases, only *g*, *h* and *i* provide values for $D(\text{CN})$ which are close to any of the possible values given in table 3. Cases *i* and *f* are unsatisfactory because, as already mentioned, the low value for $D(\text{N}_2)$ does not give a reasonable figure for the N—N bond energy in hydrazine. To get round this difficulty it would be necessary to assume that the hydrazine molecule were derived from excited nitrogen atoms. It is not possible to make an immediate decision between possibilities *g* and *h*. The spectroscopic value 170.1 kcal. for $D(\text{N}_2)$ is that which receives support from electron-impact experiments. Furthermore, the alternative figure 225 kcal. implies a high value for $D(\text{NO})$ which Flory & Johnston (1946) have criticized, mainly because it is barely to be reconciled with the results of experiments on the photo-decomposition of nitric oxide. Nor is it compatible with the appearance

potential of N^+ from nitric oxide (Hagstrum 1948). On the other hand, considerations concerning the relative energies of nitrogen bonds have led certain investigators to prefer 225 kcal. for $D(N_2)$ (Skinner 1945; Glockler 1948*b*). But in this connexion it should be noted that in his semi-empirical relationships Glockler compares other properties of carbon bonds with bond energies calculated from the 3P state of carbon.

In both cases *g* and *h*, however, there is a difficulty with respect to the CN spectrum. If $D(CN)$ is 117.9 kcal., which would follow from case *h*, the products of dissociation at the convergence limit at 60,500 cm^{-1} must be $C: ^3P + N: ^2D$. This identification of products was tentatively suggested in the appendix of an earlier paper (Long & Norrish 1946*c*), and supported on the additional grounds that the upper convergence at 65,500 cm^{-1} could be assigned the products $C: ^5S + N: ^4S$, provided the excitation energy of $C: ^5S$ were about 24,200 cm^{-1} or 69.2 kcal. This value for the $C: ^5S$ excitation energy, against which there was at the time no published experimental data, seemed to provide a confirmation of the thermochemical estimate of the energy level of the tetravalent state of carbon, namely 65 ± 10 kcal. The recent observation of the 5S level by Shenstone (1947) at 96.4 kcal., however, apart from appearing to prohibit the identification of this thermochemical estimate with the energy of the $C: ^3P \rightarrow ^5S$ transition, creates a real difficulty when 117.9 kcal. is considered for $D(CN)$, since there is no combination of atomic terms which fits the upper convergence of CN at $65,500 \pm 1000$ cm^{-1} , or lies even near to it. This argument applies with even greater force to the possibility that $D(CN)$ is 143.8 kcal. (case *g*).

The reason for this seeming incompatibility is not clear, but, in the event that $D(CN)$ has the value 117.9 kcal. (case *h*), the possibility arises that the upper $B^2\Sigma$ state of CN is derived from a hybridized tetravalent carbon atom with an internal energy around 69 kcal. above the ground state, this maintaining its identity in CN up to a high vibrational level; in spite of the fact that it does not correspond to a single atomic state. In this case the difficulty caused by the apparent convergence at 65,500 cm^{-1} would be removed. In any eventuality, some revision of the dissociation scheme of CN due to Schmid *et al.* seems to be called for, even though table 2 supports the contention of these investigators that CN does not dissociate into ground-state atoms at 172.8 kcal.

Cases *d* and *e* both presuppose an error of some 10 kcal. or more in the heat of formation derived for the CN radical, which is more than is likely. It thus appears, quite apart from the additional objections of Valatin (1948*c*), that the electron-impact value 221 kcal. for $D(CO)$ is less probable than the spectroscopic value 210.75 kcal., but it is not possible in our present state of knowledge to decide definitely against it.

CONCLUSIONS

The calculation of the heat of formation of free CN by a number of independent methods has made possible the derivation of a thermochemical relationship between $D(CO)$, $D(CN)$ and $D(N_2)$, as well as an unambiguous value for $D(NC-CN)$.

The three spectroscopic values for the dissociation energies of CO, CN and N_2 favoured by Valatin (1946*a*) are neither mutually compatible nor individually acceptable.

The spectroscopic values for $D(\text{CO})$ and $D(\text{N}_2)$ preferred by Gaydon (1947) by reason of the Non-crossing Rule are likewise mutually incompatible.

The most probable value for $D(\text{CO})$ is 210.75 kcal., corresponding to 125.0 kcal. for the heat of atomization of graphite at 0° K. This value is supported by independent lines of evidence concerning the heat of formation of methylene and its bearing on the energy involved in the stepwise dissociation of methane. The evidence examined here is decidedly incompatible with the high value 256.1 kcal. and the low value 159.7 kcal. for $D(\text{CO})$. The electron-impact value 221 kcal. appears as a less likely alternative to the spectroscopic value 210.75 kcal.

Of the three values proposed for $D(\text{N}_2)$, 115.1 kcal. appears to be definitely too low to be a possible figure. The relationship between $D(\text{CO})$, $D(\text{CN})$ and $D(\text{N}_2)$ cannot decide between 170.1 and 225.0 kcal. until $D(\text{CN})$ is definitely fixed.

From the convergence of CN at 60,500 cm.⁻¹, $D(\text{CN})$ is either 117.9 or 143.8 kcal., according to whether $D(\text{N}_2)$ is 170.1 or 225.0 kcal.

The author's grateful acknowledgments are due to Professor R. G. W. Norrish, Dr A. D. Walsh and Dr G. C. Eltenton for helpful correspondence regarding certain of the points discussed in this paper.

APPENDIX

Towards the unequivocal fixing of these important quantities, in particular the heats of atomization of carbon and nitrogen, it would be useful to close the paper by enumerating some of the ways by which the problems might be profitably approached:

(i) a precise determination of the limiting wave-length capable of effecting the photo-decomposition of carbon monoxide;

(ii) further direct measurement of the vapour pressure of graphite, including the determination of the maximum temperatures to which it is possible to heat a graphite rod electrically under varying pressures;

(iii) a calculation of the lattice energy of diamond (this would presumably provide the heat of atomization to carbon atoms in the tetravalent state);

(iv) the application of a modified Trouton's Rule to graphite;

(v) a determination of the heat of formation of diazomethane which, in conjunction with the energy absorbed in the photo-decomposition (Kirkbride & Norrish 1933), would provide an upper limit for the heat of atomization of graphite, as in the similar case of ketene;

(vi) further photochemical studies of simple cyanides, preferably in conjunction with a further analytical study of the CN spectrum;

(vii) combined photochemical and thermochemical studies of a number of simple molecules such as C_2H_2 , C_2Cl_2 , HN_3 , NH_3 , N_2H_4 and the recently discovered N_2F_2 ;

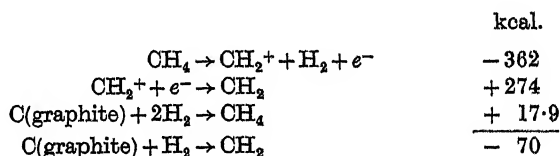
(viii) further studies of the homogeneous thermal decomposition of methane and ammonia;

(ix) determinations of the heats of reaction of processes involving CH_3 , CH_2 , CH , NH_2 , NH and CN radicals.

Note added 23 December 1948. Since the completion of this paper, other articles have appeared which have a close bearing on its contents. A revision of an earlier

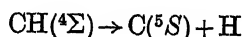
theoretical treatment of the stepwise dissociation of methane in the light of more recent experimental data has been given by Voge (1948), who concludes that the energies required for the successive removal of hydrogen atoms are $a = 101$, $b = 90$, $c = 80$ and $d = 80$ kcal., values not involving *a priori* assumptions concerning the heat of atomization of carbon. In no case do these figures differ by more than 1 or 2 kcal. from the values adopted here. This provides from a theoretical angle excellent support for the magnitude of b (~ 88 kcal.) derived entirely from experimental data in the present paper. (The alternative figures of Walsh (1947*b*) involving a triplet ground state for CH_2 , namely $a = 101$, $b = 85$, $c = 85$ and $d = 80$ kcal., do not differ greatly from those of Voge.)

Quite apart from his own considerations, Voge also cites data concerning the appearance potential of CH_2^+ from methane (Smith 1937) and the ionization energy of CH_2 (Langer & Hipple 1946), from which the heat of formation of CH_2 at room temperature may be calculated to be -70 kcal., as follows:



This corresponds to -71 kcal. at 0°K and is further experimental evidence in support of the figure ~ -70 kcal. already derived in a number of independent ways.

In direct conflict with the foregoing, a paper by Gerö (1948) on the stepwise dissociation of methane and a number of articles by Valatin (1948*a, b, d*) on allied topics extend the former contention of Schmid & Gerö (1937), already referred to here, that (with modern conversion factors) $D(\text{CO}) = 159.7$ and $L_1 = 73.9$ kcal. Bearing in mind that a and d are experimentally fixed, this would mean that in the dissociation scheme of methane the sum of b and c is more than 50 kcal. less than that adopted both in the present paper and in the paper of Voge (1948). However, by making the arbitrary and unnatural assumptions (i) that even in the CH radical carbon always remains in the tetravalent state, and (ii) that only ^5S carbon atoms and not atoms in the ^3P ground state participate in the equilibrium present in carbon vapour, the values 100.76, 97.36, 97.02 and 97.02 kcal. are obtained for the respective energies required to remove successive hydrogen atoms in methane (the second, third and fourth of these quantities being ultimately defined in a different manner from b , c and d as employed in this paper). The close agreement of the first figure with the experimental value of $D(\text{CH}_3-\text{H})$ is obtained only at the expense of further tacitly assuming (iii) that the $\text{C}-\text{H}$ bond energies in CH_3 and CH_2 groups within saturated organic molecules are identical with those in free CH_3 and free CH_2 respectively. This last assumption may involve serious errors and is in itself misleading, since it means, for example, that there remains no evidence to support the contention that the process



absorbs 97.02 kcal., as is asserted. Again, it also involves identifying bond energies with dissociation energies, for example the $\text{C}-\text{C}$ bond energy in ethane with

$D(\text{CH}_3-\text{CH}_3)$, which is not permissible (Long & Norrish 1946c). With regards assumption (i), it is hard to see why the carbon in CH, which involves only a single bonding electron from the carbon atom, should not revert to the divalent state, there being no kinetic, spectroscopic or energetic evidence to suggest that it remains tetravalent. Assumption (ii), in order to explain the discrepancy with observed equilibrium vapour-pressure measurements, involves the further improbable assumption that the 5S carbon atoms are in equilibrium with a hypothetical excited singlet state of the C_2 molecule, which cannot revert to the triplet ground state and is regarded as possessing a quadruple bond (!) whose energy is estimated to be 185 kcal. by extrapolation. However, these assumptions concerning strongly metastable atoms and molecules meet with serious objections, since the atomic lines of the $^5S \rightarrow ^3P$ transition have been observed in emission (Shenstone 1947) and the molecular lines emitted by carbon vapour inside a graphite tube below 3000° K are the ordinary triplet-triplet Swan bands (Brewer *et al.* 1948). Without these admittedly ingenious if extravagant assumptions which the Budapest school have introduced to defend a not unequivocal interpretation of the CO spectrum, there appears to be little if any independent evidence to support their interpretation. These assumptions alone are in any case not sufficient to reconcile the interpretation with all known facts, including the experimental data discussed in the present paper.

REFERENCES

- Badger, R. M. & Woo, S.-C. 1931 *J. Amer. Chem. Soc.* **53**, 2572.
Barrow, R. F., Pearson, T. G. & Purcell, R. H. 1939 *Trans. Faraday Soc.* **35**, 880.
Berthelot, M. 1875 *Ann. Chim. (Phys.)*, (v), **5**, 433.
Berthelot, M. 1881 *Ann. Chim. (Phys.)*, (v), **23**, 252.
Bichowsky, F. R. & Rossini, F. D. 1936 *The thermochemistry of the chemical substances*. New York: Reinhold.
Birge, R. T. 1941 *Rep. Progr. Phys.* **8**, 90.
Brewer, L., Gilles, P. W. & Jenkins, F. A. 1948 *J. Chem. Phys.* **16**, 797.
Büttenbender, G. & Herzberg, G. 1934 *Ann. Phys., Lpz.*, (v), **21**, 577.
Cherton, R. 1942 *Bull. Soc. Sci. Liège*, **11**, 203.
Cherton, R. 1943 *Bull. Soc. Chim. Belg.* **52**, 26.
Cutler, J. A. 1948 *J. Chem. Phys.* **16**, 136.
Eltenton, G. C. 1947 *J. Chem. Phys.* **15**, 455.
Faltings, K., Groth, W. & Harteck, P. 1938 *Z. phys. Chem. B*, **41**, 15.
Flory, P. L. & Johnston, H. L. 1946 *J. Chem. Phys.* **14**, 212.
Funke, G. W. & Lindholm, E. 1937 *Z. Phys.* **106**, 518.
Gaydon, A. G. 1947 *Dissociation energies and spectra of diatomic molecules*. London: Chapman and Hall.
Gaydon, A. G. 1948 *Nature*, **161**, 731.
Gaydon, A. G. & Penney, W. G. 1942 *Nature*, **150**, 406.
Gaydon, A. G. & Penney, W. G. 1945 *Proc. Roy. Soc. A*, **183**, 374.
Gerö, L. 1936 *Z. Phys.* **100**, 374.
Gerö, L. 1948 *J. Chem. Phys.* **16**, 1011.
Giauque, W. F. 1931 *J. Amer. Chem. Soc.* **53**, 507.
Gillespie, L. J. & Fraser, L. H. D. 1936 *J. Amer. Chem. Soc.* **58**, 2260.
Glockler, G. 1948a *J. Chem. Phys.* **16**, 600.
Glockler, G. 1948b *J. Chem. Phys.* **16**, 602.
Goldfinger, P. 1947 *Bull. Soc. Chim. Belg.* **56**, 282.
Gordon, A. R. 1937 *J. Chem. Phys.* **5**, 30.
Gordon, A. R. & Barnes, C. 1933 *J. Chem. Phys.* **1**, 692.

- Gorin, E., Kauzmann, W., Walter, J. & Eyring, H. 1939 *J. Chem. Phys.* 7, 633.
- Hagstrum, H. D. 1947 *Phys. Rev.* (ii), 72, 947.
- Hagstrum, H. D. 1948 *J. Chem. Phys.* 16, 848.
- Heim, G. 1933 *Bull. Soc. Chim. Belg.* 42, 467.
- Heimer, T. 1932 *Z. Phys.* 78, 771.
- Herzberg, G. 1939 *Molecular spectra and molecular structure. I. Diatomic molecules.* New York: Prentice-Hall.
- Herzberg, G. 1942 *J. Chem. Phys.* 10, 306.
- Herzberg, G. 1944 *Atomic spectra and atomic structure* (2nd ed.). New York: Dover.
- Herzberg, G. 1945 *Infrared and Raman spectra of polyatomic molecules.* New York: Van Nostrand.
- Herzberg, G. & Scheibe, G. 1930 *Z. phys. Chem. B*, 7, 390.
- Hogness, T. R. & Ts'ai, L.-S. 1932 *J. Amer. Chem. Soc.* 54, 123.
- Kassel, L. S. 1932 *J. Amer. Chem. Soc.* 54, 3949.
- Ketelaar, J. A. A. & Kruyer, S. 1943 *Rec. Trav. chim. Pays-Bas*, 62, 550.
- Kharasch, M. S. 1929 *Bur. Stand. J. Res., Wash.*, 2, 359.
- Kirkbride, F. W. & Norrish, R. G. W. 1933 *J. Chem. Soc.* p. 119.
- Kistiakowsky, G. B. & Gershinowitz, H. 1933 *J. Chem. Phys.* 1, 432.
- Kistiakowsky, G. B. & Van Artsdalen, E. R. 1944 *J. Chem. Phys.* 12, 469.
- Langer, A. & Hipple, J. A. 1946 *Phys. Rev.* (ii), 69, 691.
- Lemoult, P. 1909 *C.R. Acad. Sci., Paris*, 148, 1602.
- Linnett, J. W. 1940 *J. Chem. Phys.* 8, 91.
- Linnett, J. W. 1941 *Trans. Faraday Soc.* 37, 469.
- Long, L. H. 1948 *J. Chem. Phys.* 16, 1087.
- Long, L. H. & Norrish, R. G. W. 1946a *Nature*, 157, 486.
- Long, L. H. & Norrish, R. G. W. 1946b *Nature*, 158, 237.
- Long, L. H. & Norrish, R. G. W. 1946c *Proc. Roy. Soc. A*, 187, 337.
- Long, L. H. & Walsh, A. D. 1947 *Trans. Faraday Soc.* 43, 342.
- Mooney, R. B. & Reid, H. G. 1931 *Nature*, 128, 271.
- Mooney, R. B. & Reid, H. G. 1932 *Proc. Roy. Soc. Edinb.* 52, 152.
- Neuimin, H. H. & Terenin, A. N. 1936 *Acta physicochim. U.R.S.S.* 5, 465.
- Norrish, R. G. W. 1934 *Trans. Faraday Soc.* 30, 103.
- Norrish, R. G. W., Crone, H. G. & Saltmarsh, O. 1933 *J. Chem. Soc.* p. 1533.
- Paneth, F. A., Hofeditz, W. & Wunsch, A. 1935 *J. Chem. Soc.* p. 372.
- Pearson, T. G., Purcell, R. H. & Saigh, G. S. 1938 *J. Chem. Soc.* p. 409.
- Prosen, E. J., Pitzer, K. S. & Rossini, F. D. 1945 *Bur. Stand. J. Res., Wash.*, 34, 403.
- Rice, F. O. & Glasebrook, A. L. 1934 *J. Amer. Chem. Soc.* 56, 2381.
- Rice, F. O. & Greenberg, J. 1934 *J. Amer. Chem. Soc.* 56, 2268.
- Robertson, N. C. & Pease, R. N. 1942a *J. Chem. Phys.* 10, 490.
- Robertson, N. C. & Pease, R. N. 1942b *J. Amer. Chem. Soc.* 64, 1880.
- Rosenblum, C. 1938 *J. Amer. Chem. Soc.* 60, 2819.
- Rosenblum, C. 1941 *J. Amer. Chem. Soc.* 63, 3322.
- Ross, W. F. & Kistiakowsky, G. B. 1934 *J. Amer. Chem. Soc.* 56, 1112.
- Rossini, F. D., Wagman, D. D., Evans, W. H., Blau, E. J. & Levine, S. 1947 *Selected values of chemical thermodynamic properties.* Washington: National Bureau of Standards.
- Schmid, R. F. & Gerö, L. 1935 *Z. Phys.* 93, 656.
- Schmid, R. F. & Gerö, L. 1937 *Z. phys. Chem. B*, 36, 105.
- Schmid, R. F. & Gerö, L. 1943 *Matematikai és Természettudományi Értesítő*, 62, 416.
- Schmid, R. F. & Gerö, L. 1946 *Proc. Phys. Soc.* 58, 701.
- Schmid, R. F., Gerö, L. & Zemplén, J. 1938 *Proc. Phys. Soc.* 50, 283.
- Semerano, G. & Riccoboni, L. 1941 *Z. phys. Chem. A*, 189, 203.
- Shenstone, A. G. 1947 *Phys. Rev.* (ii), 72, 411.
- Shidei, T. 1936 *Jap. J. Phys.* 11, 23.
- Skinner, H. A. 1945 *Trans. Faraday Soc.* 41, 645.
- Smith, L. G. 1937 *Phys. Rev.* (ii), 51, 263.
- Springall, H. D. 1947 *Trans. Faraday Soc.* 43, 177.
- Steacie, E. W. R. 1946 *Atomic and free radical reactions.* New York: Reinhold.
- Stevenson, D. P. 1939 *J. Chem. Phys.* 7, 171.
- Terenin, A. N. & Neuimin, H. H. 1934 *Nature*, 134, 255.

- Terenin, A. N. & Neušmin, H. H. 1935 *J. Chem. Phys.* 3, 436.
 Thompson, H. W. 1941 *Trans. Faraday Soc.* 37, 344.
 Thomsen, J. 1905 *Z. phys. Chem.* 52, 343.
 Trenner, N. R., Morikawa, K. & Taylor, H. S. 1937 *J. Chem. Phys.* 5, 203.
 Valatin, J. G. 1946a *J. Chem. Phys.* 14, 568.
 Valatin, J. G. 1946b *Proc. Phys. Soc.* 58, 695.
 Valatin, J. G. 1948a *Rev. sci., Paris*, 86, 135.
 Valatin, J. G. 1948b *J. chim. phys.* 45, 123.
 Valatin, J. G. 1948c *Phys. Rev.* (ii), 74, 350.
 Valatin, J. G. 1948d *J. Chem. Phys.* 16, 1018.
 Voge, H. H. 1936 *J. Chem. Phys.* 4, 581.
 Voge, H. H. 1948 *J. Chem. Phys.* 16, 984.
 Wagman, D. D., Kilpatrick, J. E., Taylor, W. J., Pitzer, K. S. & Rossini, F. D. 1945 *Bur. Stand. J. Res., Wash.*, 34, 143.
 Walsh, A. D. 1947a *Trans. Faraday Soc.* 43, 60.
 Walsh, A. D. 1947b *Faraday Soc. Discussions*, 2, 18.
 Walsh, A. D. 1948 *J. Chem. Soc.* p. 398.
 White, J. U. 1940 *J. Chem. Phys.* 8, 502.
 Wicke, E. 1942 *Ergebn. exakt. Naturw.* 20, 1.
 Wicke, E. 1945 *Die Chemie*, 58, 16.
 Yakovleva, A. V. 1938 *Acta physicochim. U.R.S.S.* 9, 665.

The behaviour of waves on tidal streams

BY N. F. BARBER, *The Admiralty Research Laboratory, Teddington, Middx.*

(Communicated by G. E. R. Deacon, F.R.S.—Received 4 November 1948)

The paper discusses the manner in which waves change their characteristics when they pass through regions where the water has a streaming motion. The treatment applies to tidal streams, the velocity of which depends both on time and position. Some experimental evidence is provided in support of the theory.

INTRODUCTION

Since early in 1945 wave records from the coast of Cornwall have been submitted to frequency analysis for the purpose of studying the generation and propagation of waves and swell. Some results have been reported by Barber & Ursell (1948). The evidence indicates that in a storm, trains of waves are generated of a variety of wave-lengths up to a maximum wave-length depending on the wind strength and that each wave train advances across the ocean with a speed approximately equal to $gT/4\pi$, which is the group velocity indicated by theory for waves whose period is T . The separation between the longest and shortest waves travelling in the same direction must therefore increase with the distance travelled from the generating area, and because of this dispersion the first wave trains to arrive at a distant recording station will exhibit a long natural period. Wave trains of shorter period arrive later, and it is to be expected that the swell arriving at the coast will show a period which decreases continuously with time.

The frequency spectra of waves recorded at Pendeen and Perranporth illustrate this behaviour, but they show that the period of the swell arriving at the coast does not vary exactly in accordance with the simple theory. The swell from a distant storm is observed to have a period which fluctuates by as much as ± 1 sec. in cycles of $12\frac{1}{2}$ hr. about the smooth curve drawn through the observations to represent the general trend towards shorter periods. This fluctuation is evident in the curves of figure 1, which are based on observations made about the time of spring tides.

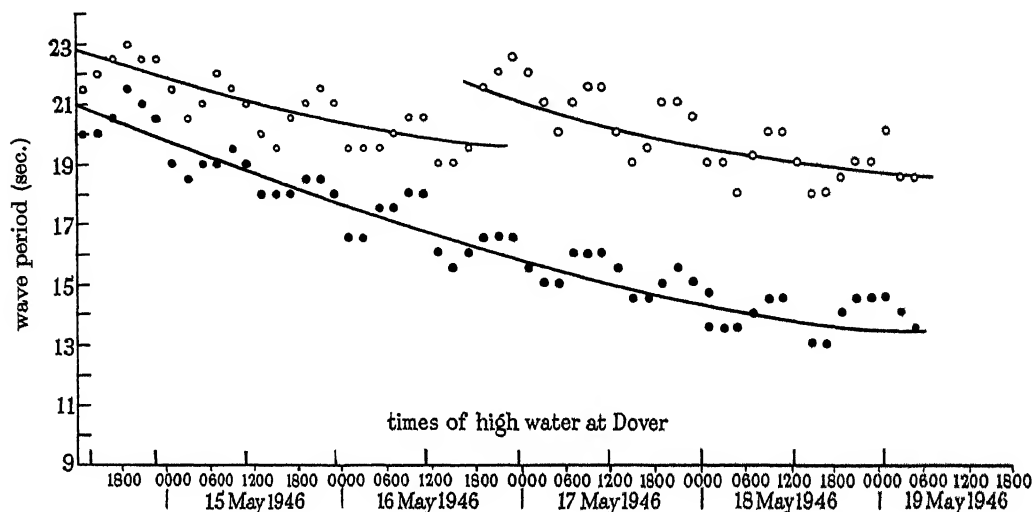


FIGURE 1. The maximum and minimum periods limiting the 24 to 14 sec. frequency band in the wave spectra of 14 to 19 May 1945 at Perranporth.

The fluctuation in period has been attributed to the effect of tidal streams through which the swell had to travel in the last 200 miles of its journey to the coast. The present aim is to discuss this effect in more detail.

THEORY OF THE STEADY STATE

Unna (1942) has discussed the steady state of a wave train entering streams whose velocity does not change with time. He shows that where the stream has a velocity u the waves may be expected to exhibit a new wave-length λ and velocity c given by the rule

$$\lambda_0/c_0 = \lambda/(c+u),$$

where the zero subscript denotes the values of these quantities for that part of the wave train which is on slack water. The equality is obtained by assuming that an equal number of wave crests per unit time must pass any fixed observer. The argument presumes that it is physically possible to have a train of waves, extending from slack water to moving water, whose waves retain their identity, no new ones appearing and none disappearing. No description of such a train has yet been

given in terms of a velocity potential, and this may suggest that such a train cannot exist; there may, for instance, be some reflexion of the wave train. But if the existence of the wave train is assumed, and if the wave-length and wave velocity of any part are related by the usual equation, the changes in velocity or length may be predicted; in deep water the rule

$$\lambda = 2\pi c^2/g$$

enables the first equation to be written in the form

$$0 = (c/c_0)^2 - c/c_0 - u/c_0,$$

which is equivalent to formulae due to Unna. The equation describes two types of wave train:

(1) If c/c_0 is greater than $\frac{1}{2}$, the wave train is one which may extend into slack water; it cannot extend into an opposing stream whose velocity is greater than $\frac{1}{2}c_0$. Before this point the waves steepen and break.

(2) If c/c_0 is less than $\frac{1}{2}$, the wave train is one which rides on an opposing stream whose velocity is greater than the group velocity, $\frac{1}{2}c$, of the waves. The train cannot extend into slack water; before it arrives there the waves steepen and break. This wave train might be generated as the wake of a boat moving up stream.

The arguments have been extended by Johnson (1947) to the two-dimensional problem of waves crossing obliquely the boundary between two currents. He infers that when the waves approach the boundary obliquely they may fail to cross it, their energy being dissipated by reflexion or in breaking.

THE NON-STEADY PROBLEM

It will be appreciated that the tidal variations of period observed at Pendeen and Perranporth cannot appropriately be dealt with by steady-state theory. The velocities of the tidal streams change considerably during the time that the swell takes to cross the area of streams lying between the deep sea and the coast. It is desirable to develop some treatment suitable for waves on streams whose velocity changes with time.

Unna (1941), discussing the behaviour of short waves riding on swell, treats the short waves as if they expanded and contracted with the water mass on which they move. The short waves then attain their greatest length as the trough of swell passes them and attain their smallest wave-length when they are overtaken by the crest. This somewhat intuitive treatment can be justified by reducing the problem to the steady state. A uniform velocity impressed on the system will bring the swell to rest and it may be regarded as slight variations in space, but not in time, of a rapid stream on which the short waves move. This leads to the formulae given by Unna.

The idea that waves expand or contract with the water mass can be justified, however, by another means which is found to provide a solution to the general problem. This treatment is given below, and its conclusions will be seen to be in fairly good agreement with experimental observations.

KINEMATICS OF WAVES ON STREAMS

The classical theory of water waves shows that if the height of the waves is sufficiently small the system can always be regarded as due to the superposition of a sufficient number of elementary wave trains, each consisting of long parallel crests equally spaced and moving with a speed appropriate to the wave-length. In the resulting interference pattern the elementary waves lose their identity, and the more complicated phenomena of group motion, in which waves appear and disappear as they travel in groups, take their place.

In discussing waves on streaming water two assumptions will be made. By analogy with the theory of waves on slack water it will be assumed that a complicated wave motion on streaming water can be looked upon as being the superposition of a sufficient number of elementary trains which could have an independent existence and in which the wave crests would retain their existence as they progressed through the streaming water. Whether such elementary wave trains are possible on streaming water is not known, but the analogy with classical wave theory makes it plausible.

It will also be assumed that in each of these elementary trains of waves the length of the wave and its velocity relative to the water are related by the equation of classical theory

$$c^2 = g \frac{\lambda}{2\pi} \tanh \frac{2\pi h}{\lambda}, \quad (1)$$

where c = wave velocity relative to the water mass, λ = wave-length, g = acceleration due to gravity, h = water depth.

This equation is strictly applicable only to the waves of an elementary train in slack water, but it seems likely that it may apply to an elementary train on streaming water provided:

- (a) that the changes in stream velocity are small during a very large number of wave periods;
- (b) that the stream velocity is very nearly constant over a large number of wave-lengths;
- (c) that the streaming velocity is sensibly uniform over a depth equal to half a wave-length.

With these assumptions it is possible to consider the one-dimensional case illustrated in figure 2. The velocity u of the stream is prescribed at each point x as a function of the time t . An elementary train of waves is present on the surface, and the length λ and velocity c of the waves will in general be functions of both the position x and time t .

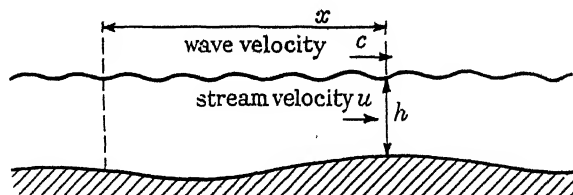


FIGURE 2. Kinematics of waves on streams (one-dimensional problem).

Considering a wave crest which is at position x at time t , its velocity relative to the water is c and relative to the fixed system of co-ordinates is $(c+u)$. The N th wave in succession from this one will be at a distance $N\lambda$ from it, and the velocity of this wave will differ slightly from the velocity of the first one, the difference being

$$N\lambda\partial(c+u)/\partial x.$$

This relative velocity may also be expressed as the time rate of change of the distance between the two crests. The value of λ depends both upon x and t , and since the waves move forward a distance $(c+u)\delta t$ in a brief interval δt , the rate of change of $N\lambda$ with time is

$$N(\partial\lambda/\partial t + (c+u)\partial\lambda/\partial x).$$

The two expressions for relative velocity may be equated to give the relation

$$\frac{1}{\lambda}\left(\frac{\partial\lambda}{\partial t} + (c+u)\frac{\partial\lambda}{\partial x}\right) = \frac{\partial(c+u)}{\partial x}. \quad (2)$$

WAVES ON WATER WHICH IS DEEP OR OF CONSTANT DEPTH

The formulae (1) and (2) may now be combined, but the precise form of equation (1) is not involved at this stage, and it is only necessary to observe that the velocity c has been assumed to be a function of λ , g and h and not a function, for instance, of the space and time derivatives of these quantities.

In the first instance it will be assumed that the waves are in water whose depth is uniform and constant, or else that the water in which the waves are travelling is deeper than a wave-length. The equation (1) then shows that we may write

$$\frac{\partial c}{\partial x} = \frac{dc}{d\lambda} \frac{\partial\lambda}{\partial x}.$$

This relation when substituted in equation (2) gives

$$\frac{1}{\lambda}\left[\frac{\partial\lambda}{\partial t} + \left(u + c - \lambda \frac{dc}{d\lambda}\right) \frac{\partial\lambda}{\partial x}\right] = \frac{\partial u}{\partial x}. \quad (3)$$

The right-hand side of this equation measures the rate of expansion of the water surface which is produced by the streaming motion. The left-hand side is the fractional rate of increase of wave-length with time; the expression does not refer to a given set of waves, but to those waves which happen to be in the vicinity of an observer who moves relatively to the water with a velocity

$$c - \lambda dc/d\lambda.$$

This expression is the usual definition of the group velocity of the waves in classical theory. In this problem, therefore, as in many others, the group velocity enters into the equation with a peculiar significance. It appears that if an observer follows a given wave group through moving water, *the average length of the waves in the group expands or contracts at the same rate as the general surface of the water upon which the group is moving.* It will be appreciated that this result determines

the whole behaviour of the waves, for the velocity and period may be inferred when the wave-length has been found.

It is convenient to use a special symbol to denote the time rate of change of some characteristic of the waves in a particular group, and using the notation

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \left(u + c - \lambda \frac{\partial c}{\partial \lambda} \right) \frac{\partial}{\partial x}, \quad (4)$$

the equation (3) may be written as

$$\frac{1}{\lambda} \frac{D\lambda}{Dt} = \frac{\partial u}{\partial x}. \quad (5)$$

THE GENERAL CASE

It is clear that in tidal regions the water depth is a function both of place and time, and the effective value of gravity to be used in equation (1) may also vary; the waves are riding on a mass of water whose vertical acceleration must modify the effective force of gravity on the waves. It can be shown that these two effects rarely appear simultaneously; the changes in gravity are of importance in discussing the behaviour of wind waves riding on long swell in water whose depth is comparable with the length of the swell, and the changes in depth are of importance in discussing the behaviour of waves on tides. For the sake of generality, changes in both h and g will be discussed.

The wave equation (1) is a relation between the quantities c , λ , g and h , and if the wave period is denoted by T , where

$$T = \lambda/c,$$

it may be shown that the following relations hold between the partial differential coefficients

$$\left. \begin{aligned} \left(\frac{\partial c}{\partial h} \right)_{\lambda g} &= \frac{1}{\lambda} \left(\frac{\partial \lambda}{\partial h} \right)_{xg} \left(c - \lambda \left(\frac{\partial c}{\partial \lambda} \right)_{gh} \right), \\ \left(\frac{\partial c}{\partial g} \right)_{\lambda h} &= \frac{1}{\lambda} \left(\frac{\partial \lambda}{\partial g} \right)_{xh} \left(c - \lambda \left(\frac{\partial c}{\partial \lambda} \right)_{gh} \right), \end{aligned} \right\} \quad (6)$$

where the subscripts denote the quantities held constant during the differentiation. The quantity $c - \lambda(\partial c/\partial \lambda)_{gh}$ in these expressions is recognizable as the group velocity and will be denoted by G .

Substitution may now be made in the kinematic relation (2). The quantity $\partial c/\partial x$ on the right-hand side of this equation may be expanded as

$$\frac{\partial c}{\partial x} = \left(\frac{\partial c}{\partial \lambda} \right)_{gh} \frac{\partial \lambda}{\partial x} + \left(\frac{\partial c}{\partial h} \right)_{\lambda g} \frac{\partial h}{\partial x} + \left(\frac{\partial c}{\partial g} \right)_{\lambda h} \frac{\partial g}{\partial x},$$

and using the equations (6) the kinematic equation becomes

$$\frac{1}{\lambda} \frac{D\lambda}{Dt} = \frac{\partial u}{\partial x} + \frac{1}{\lambda} \left(\frac{\partial \lambda}{\partial h} \right)_{xg} G \frac{\partial h}{\partial x} + \frac{1}{\lambda} \left(\frac{\partial \lambda}{\partial g} \right)_{xh} G \frac{\partial g}{\partial x}. \quad (7)$$

This expression for the fractional rate of increase of wave-length in a wave group is similar to the equation (5) derived previously, but it includes two terms

which show the effects of changes in depth and in effective gravity. This may be regarded as the fundamental equation governing the change of character of the waves of a group as the group advances through streaming water. From the wave-length the period or velocity or group velocity may be inferred, but it is possible to obtain the changes in these quantities explicitly. Thus if M denotes one of the quantities c , T or G it can be shown that

$$\begin{aligned} \frac{DM}{Dt} = & \left(\frac{\partial M}{\partial \lambda} \right)_{gh} \lambda \frac{\partial u}{\partial x} + \left(\frac{\partial M}{\partial h} \right)_{g\lambda} \left(\frac{\partial h}{\partial t} + u \frac{\partial h}{\partial x} \right) + \left(\frac{\partial M}{\partial h} \right)_{Tg} G \frac{\partial h}{\partial x} \\ & + \left(\frac{\partial M}{\partial g} \right)_{h\lambda} \left(\frac{\partial g}{\partial t} + u \frac{\partial g}{\partial x} \right) + \left(\frac{\partial M}{\partial g} \right)_{Th} G \frac{\partial g}{\partial x}. \end{aligned} \quad (8)$$

A reasonable physical interpretation may be given to equations (7) and (8). From equation (7) it will be seen that the time derivatives of g and h do not affect the wave-length, and so long as the waves remain on a mass of water which is all behaving in a very similar way the changes in wave-length are produced solely by the contraction or expansion of the water surface; changes in g or h only affect the wave-length in so far as the group experiences them as a result of moving with velocity G on to new masses of water where g or h are different. When moving on to new water, the changes in wave-length occasioned by changes in g or h take place as if the period of the waves remained constant. Thus if waves are present on the surface of water in a tank which suddenly begins to accelerate upwards or downwards the length of the waves will not change, but the period will alter appropriately to the new value of g ; when waves pass into shoaling water which has no streaming motion the period remains constant, but the wave-length alters appropriately to the new depth of water. The other wave characters experience changes which may be inferred from these changes in wave-length or period.

CHANGES IN THE APPARENT PERIOD

A wave characteristic which is not included in the general treatment of the previous section is the apparent period of the waves as recorded by a stationary observer. This will be written as

$$T^* = \lambda / (c + u), \quad (9)$$

and it is this apparent period which is measured in the frequency spectra of waves recorded by an instrument fixed upon the sea bed.

The differentiation of equation (9) gives

$$\frac{1}{T^*} \frac{DT^*}{Dt} = \frac{1}{\lambda} \frac{D\lambda}{Dt} - \frac{1}{(c+u)} \left(\frac{Dc}{Dt} + \frac{Du}{Dt} \right),$$

where the operator D/Dt has the significance previously given to it in (4). The quantities $D\lambda/Dt$, Dc/Dt and Du/Dt may be obtained from equations (7), (8) and (4), and upon substitution of these quantities the equation governing the apparent period appears as

$$\frac{1}{T^*} \frac{DT^*}{Dt} = - \frac{1}{(u+c)} \left[\frac{\partial u}{\partial t} + \left(\frac{\partial c}{\partial h} \right)_{\lambda g} \frac{\partial h}{\partial t} + \left(\frac{\partial c}{\partial g} \right)_{\lambda h} \frac{\partial g}{\partial t} \right]. \quad (10)$$

This equation shows that the changes in the apparent period depend only upon the time derivatives of u , g or h ; if, therefore, a wave group passes through an area of sea in which the streaming motion is everywhere constant with time, the apparent period of the waves in the group remains unchanged throughout its progress. The equation shows how the apparent period will change if the waves pass through a tidal area in which u , g and h vary with time.

When wave groups complete their passage through a tidal area in a small fraction of a tidal cycle the change in apparent period will be proportionately small. Comparing two wave-measuring stations screened from the ocean by different widths of continental shelf, it may be inferred that the corrections which must be made to the measured periods in order to obtain the true wave periods in the open sea will be smaller for the station with the narrower continental shelf, assuming that the stream velocities and tidal ranges are similar in the two cases. The proposed wave measuring stations at Casablanca and Wellington, N.Z. are likely therefore to provide a simpler picture of the state of the sea than do the stations in Cornwall.

A PRACTICAL INSTANCE OF THE TIDAL VARIATION IN APPARENT PERIOD

(a) *Swell travelling from the Southern Ocean*

The equations (5), (7) and (8) derived in the previous sections make it possible to deduce what changes will take place in waves when they cross an area of sea in which the stream velocity and the depth of water are known as functions of time and position. In any practical case it is unlikely that the stream velocity and depth can be expressed as analytic functions of x and t , and a graphical method is more suitable for the integration of the equations. This method has been used to estimate the total variation to be expected in the apparent period of a wave train having a period of 18 sec. in the open sea, which arrives at Perranporth, Cornwall, after being generated in the region of the Falkland Islands. The fluctuations observed experimentally have been shown in figure 1.

Figure 3 is a refraction diagram for such waves as they cross the continental shelf, and the full line represents the path of a wave group arriving at Perranporth. Figure 4 shows the velocity of the tidal streams at various positions on this path throughout the tidal cycle, the value used being the component of the stream velocity parallel to the line of the path; these values have been estimated from the Admiralty Atlas of tidal streams. Most of the curves of figure 4 are approximately sinusoidal, and irregularities are probably due to errors in interpolating from the Atlas. They are more complicated in the region of Perranporth, for the tidal streams on the north Cornish coast appear to show a double maximum flood stream. Figure 5 shows the mean depth of water at points along the path, values of the tidal range in depth and the time interval by which high water at various points along the path precedes high water at Dover; this information is obtained from the Admiralty Atlas of tides, and high water at Dover is used as a reference time so as to conform with usual practice.

The data of figures 4 and 5 were used in the integration of equation (10) to obtain the total change in apparent period which the waves would show at Perranporth; only the first two terms on the right-hand side of this equation were considered. The values of the quantity $\partial u/\partial t$ were obtained for points at intervals of 10 miles along the path by taking differences between successive hourly values of the stream velocities in figure 4. These values were written in array upon a diagram whose axes were the distance from Perranporth and the time relative to high water at Dover. On this diagram straight lines were drawn to represent the

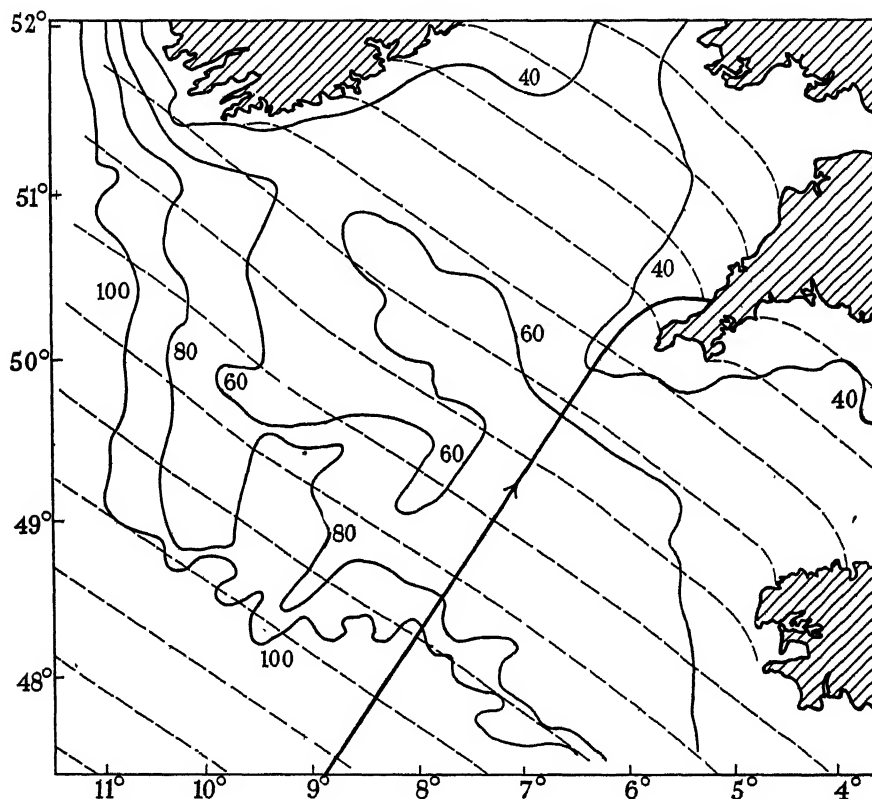
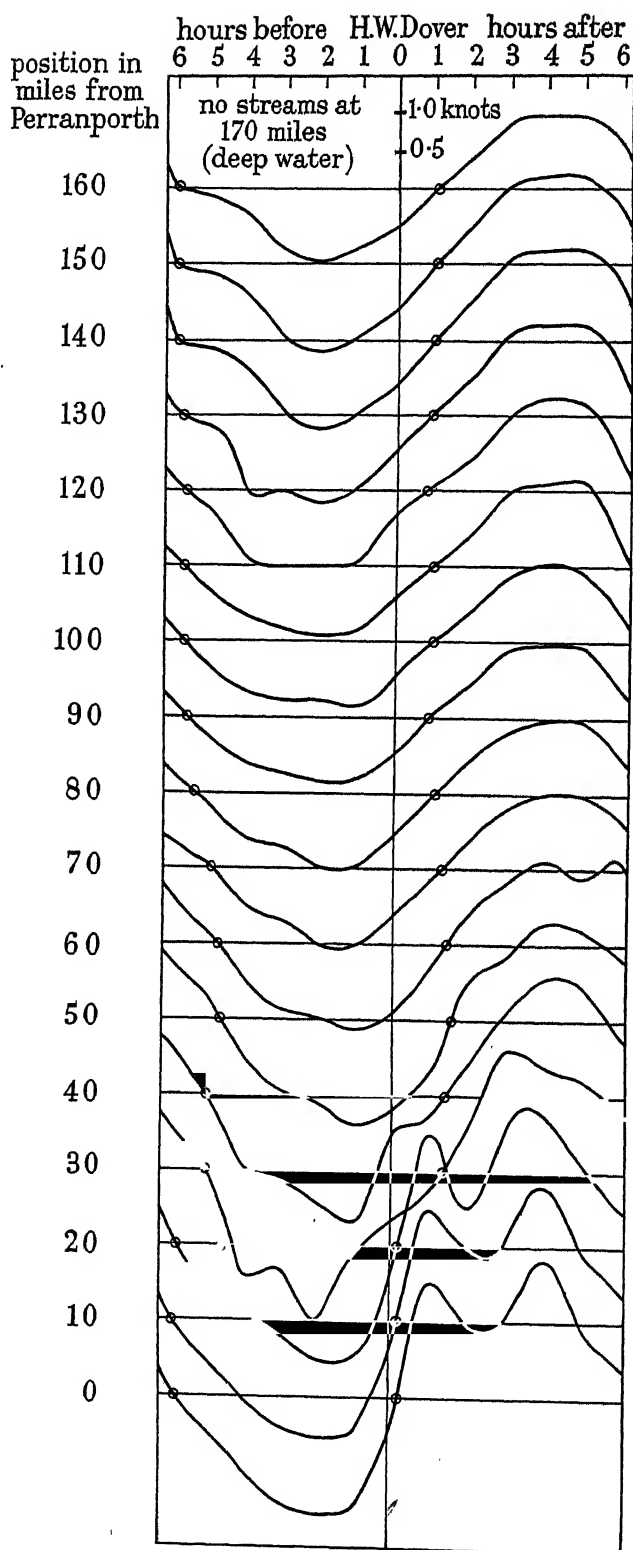


FIGURE 3. Refraction diagram for waves crossing the continental shelf from the direction 215°. Depth contours in fathoms. Broken lines show every 100th wave crest. Full line shows the path of a group arriving at Perranporth.

progress of a wave group arriving at Perranporth at various hours before and after the time of high water at Dover, the gradient of the lines being appropriate to a uniform velocity of the group equal to 32 knots. Some approximation is involved here, since the different depths of water along the path lead to theoretical values of the group velocity ranging from 33.5 to 30.0 knots, and the tidal streams themselves are of the order of a knot and affect the progress of the group over ground. These differences mean that the position of a wave group at hourly intervals would be somewhat different from the positions obtained by the assumption of a constant value for the group velocity, and the values of $\partial u/\partial t$ at these positions



Graphs against time, of the component of velocity of tidal stream (at springs) in the direction of wave progress from Falkland Islands, past Land's End to Perranporth. The curves are based on the Admiralty charts of tidal streams issued for hourly intervals relative to high water at Dover. The horizontal lines are at intervals of 1 knot on the vertical scale. The times at which the streams are zero are indicated by small circles.

FIGURE 4. Variations of the component of tidal stream at various points upon the path in figure 3. The velocity plotted is the component of stream velocity along the path.

would differ in consequence; it was considered, however, that assumption of a constant group velocity of 32 knots would not lead to any serious error. From the diagram it was possible to infer the values of $\partial u/\partial t$ which the groups would experience at equal intervals during their progress and to obtain, by addition, the

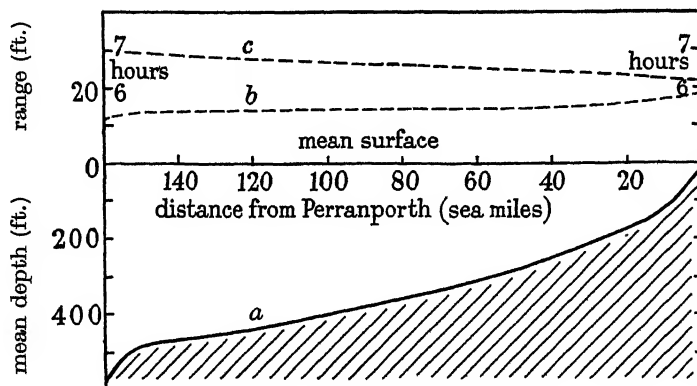


FIGURE 5. Mean depth of water at various points along the path shown in figure 3. *a*, mean depth of water; *b*, range of tide at springs; *c*, hours by which high tide precedes the time of high tide at Dover.

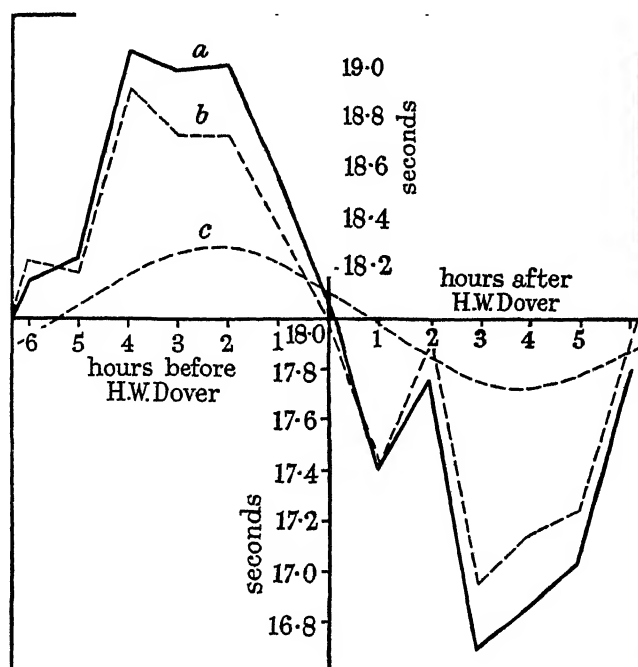


FIGURE 6. Estimated values of apparent period of waves originally of 18 sec. period which travel across the continental shelf in a direction 035° as in figure 3, and arrive at Perranporth at various tidal times. *a*, total value; *b*, effect of tidal stream velocities; *c*, effect of changing water depth in tides.

integrated value of the first term in equation (10). The changes in apparent period, due to the term $\partial u/\partial t$, which the various groups would exhibit on arrival at Perranporth are shown by the broken line *b* in figure 6. It seems likely that much of the erratic nature of this curve is due to errors in interpolation in reading from the Atlas. The large fluctuation at 2 hr. after high water at Dover may be real however, since it is associated with the double flood stream near the north Cornish coast.

The integral of the second term in equation (10) was obtained somewhat differently. The maximum value of the quantity $\partial h/\partial t$ at various positions on the path was obtained from the tidal ranges shown in figure 4 on the assumption that the rise and fall of water-level was sinusoidal with tidal period, and the values of the coefficient $\frac{1}{c} \left(\frac{\partial c}{\partial h} \right)_{\lambda_0}$ were evaluated for the various positions on the path using the mean water depths shown in figure 5 and assuming a value of 18 sec. for the wave period. The quantities were then compounded in a vector diagram, the amplitudes of the vector elements being the maximum values of the quantity

$$\frac{1}{c} \left(\frac{\partial c}{\partial h} \right)_{\lambda_0} \frac{\partial h}{\partial t}$$

obtaining at the various positions on the path, and the angles of the vector elements being the hour angle by which a wave group arriving at Perranporth at the time of high water at Dover would pass each of the positions prior to the time of maximum $\partial h/\partial t$ at that position. The summation of these elements in the vector diagram showed that the greatest increase in apparent period due to this cause would occur in waves arriving at Perranporth $2\frac{1}{4}$ hr. before high water at Dover, and it would amount to an increase of 0.28 sec.; for the remainder of the tidal cycle it is indicated by the sinusoidal broken curve *c* in figure 5.

The sum of the two calculated changes is shown as the full line *a* in figure 6. Disregarding certain irregularities in the curve it appears that the overall fluctuation in apparent period should be about 2.3 sec., and the greatest period should be shown by waves arriving at Perranporth about 3 hr. before the time of high water at Dover. This is in fair agreement with the fluctuations in the apparent period evident in figure 1.

(b) *Swell travelling from the west*

Calculations similar to those above have been made to find the change in apparent period that might be expected in swell arriving at Perranporth after having crossed the continental shelf from the west. The calculated curves are shown in figure 7; these suggest that the greatest increase in period would be about 0.6 sec. and would be shown by swell arriving at Perranporth about half an hour before the time of high water at Dover. The swell generated in storms in the North Atlantic does not usually show such regular and well-defined fluctuations in period as does swell coming from the Southern Ocean, and it can only be said that the calculated changes shown in figure 7 are not in violent disagreement with the observations.

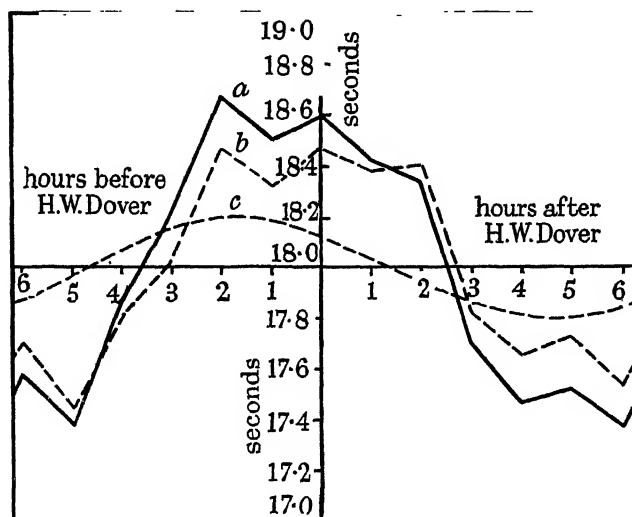


FIGURE 7. Estimated values of apparent period of waves originally of 18 sec. period which travel across the continental shelf in a direction 090° and arrive at Perranporth at various tidal times. *a*, total value; *b*, effect of tidal streams; *c*, effect of changing water depth in tides.

CONCLUSIONS

Fluctuations in tidal cycles of the period of waves recorded by a stationary instrument can satisfactorily be attributed to the action of tidal streams. In deep water the average length of waves appears to expand or contract at the same rate as the general surface of water on which they are moving. The change in period will be proportionately small when the tidal streams are weak or where the waves complete their passage through the tidal area in a small fraction of a tidal cycle.

The author wishes to express his indebtedness to P. J. H. Unna for a number of suggestions on this and on allied subjects. The data of figure 1 are based on frequency analyses made by J. Darbyshire. The paper is published by permission of the Chief of the Royal Naval Scientific Service.

REFERENCES

- Barber, N. F. & Ursell, F. 1948 *Phil. Trans. A*, **240**, 527.
- Johnson, J. W. 1947 *Trans. Amer. Geophys. Un.* **28**, 867.
- Unna, P. J. H. 1941 *Nature*, **148**, 226.
- Unna, P. J. H. 1942 *Nature*, **149**, 219.
- Unna, P. J. H. 1947 *Nature*, **159**, 239.

Kinetic theory of diffusion in gases and liquids

I. Diffusion and the Brownian motion

BY L. M. YANG, *Department of Mathematical Physics, University of Edinburgh*

(Communicated by M. Born, F.R.S.—Received 25 November 1948—

Revised 10 February 1949)

In the present paper the phenomenon of diffusion is examined in the light of the theory of the Brownian motion. The coefficients of self-diffusion, ordinary diffusion and thermal diffusion are expressed in terms of the first and second moments of certain transition probabilities familiar in the theory of the Brownian motion. It is then found possible in gases of low or moderate density where a fairly well-defined free path exists to follow the future course of a given molecule statistically to as many free flights as required provided the velocity distribution of the molecules in the medium is known. This consideration on the one hand leads to a rigorous expression for the coefficient of self-diffusion directly calculated from a Maxwellian distribution, and on the other serves to clarify the relation between the old free-path theory of gaseous diffusion and the rigorous theory of gaseous diffusion and between self-diffusion and mutual diffusion. Further, an approximate theory of diffusion in liquids corresponding to the old free-path theory in gases is suggested.

1. INTRODUCTION

The kinetic theory of diffusion in gases has been approached in two different ways in the past. The first was by means of the free-path theory which, though unable to give exact quantitative results owing to the assumption that molecules after collision are distributed according to the Maxwellian velocity distribution in non-uniform gases, provides a simple, vivid picture of the complex molecular transport phenomena in gases (Jeans 1925). The second one due to Chapman and Enskog was to determine the deviation from the Maxwellian velocity distribution in a non-uniform gas by solving Maxwell and Boltzmann's collision equation (Chapman & Cowling 1939). This has been successful in most respects for gases at low density, but the generalization for the condensed phase leads to almost insurmountable difficulties. In the present paper the phenomenon of diffusion is examined in the light of the theory of Brownian motion (Chandrasekhar 1943). It is found that the coefficient of self-diffusion in a simple gas or a gas mixture can be rigorously calculated by treating the gas molecules as Brownian particles. Consistent with the assumption of a constant coefficient of diffusion it is shown in the theory of the Brownian motion that the mean square displacement is proportional to the time interval provided that the time interval is small, so that the mean square displacement is macroscopically small but at the same time large enough for the particle to suffer a large number of collisions with the surrounding molecules. The constant ratio is obviously a physical property of the medium in which the particle moves. The relation of this constant ratio to molecular data, which is left untouched in the theory of Brownian motion, can be obtained only if one can follow the motion of particle within the time interval described above. By taking advantage of the simplified features of a gas, it has been found possible to formulate rigorously the probability of a series of successive flights

and to follow statistically the future course of a given molecule for as many successive flights as one requires. Thereby the above-mentioned ratio is explicitly expressed in terms of molecular data and is shown to approach a plateau value as the number of free flights occurring in the interval becomes large. The present method of calculating the coefficient of self-diffusion has the advantage that it requires only the equilibrium velocity distribution and is easily extended to apply to mixtures of more than two components. The same method applied to a non-uniform gas brings out the essentially different nature of self-diffusion in uniform fluids and mutual diffusion in non-uniform fluids, reveals the nature of approximation inherent in the free-path theory, and enhances the necessity for obtaining exactly the local true velocity distribution, consistent with the existing non-uniform state of the fluid.

In spite of the different nature of self-diffusion and mutual diffusion mentioned above, the latter can always be expressed approximately in terms of the former; the old free-path theory is just a simple example here. A simple explanation of the phenomena of thermal diffusion is here obtained. It also suggests that in the liquid region a corresponding approximate theory is possible, provided that the coefficient of self-diffusion in a uniform liquid mixture is known. The strict theory of diffusion in liquids, however, will require finding the true local velocity distribution. This will be undertaken in part II.

2. DIFFUSION AND TRANSITION PROBABILITY

The transport of molecular properties in a non-uniform fluid proceeds in two different ways: the first is by means of the migration of molecules from one place to another carrying with them the local properties—the densities, the mass velocity and the temperature—this may be called the kinetic part of the transport process; the second way, called the potential part of the transport process, is due to the action of intermolecular forces. Diffusion is decidedly the simplest of all transport processes, as it involves only the transport of number densities. The process of diffusion in a mixture can be described in all cases by means of a transition probability $\psi_i(\mathbf{x}, \Delta\mathbf{x}, t, \Delta t)$ defined as the probability that a molecule of the i th kind in the mixture at \mathbf{x} and at time t will suffer a displacement $\Delta\mathbf{x}$ in time Δt . ψ_i depends in general on \mathbf{x} and t through the local parameters λ_i ($\lambda_i = n_i, \mathbf{u}, T, E_i$) and their space gradients of all order where n_i is the number density of the i th kind of molecules, T the temperature, \mathbf{u} the mass motion velocity and E_i the potential energy of the i th kind of molecule in the external force field. The time interval Δt is supposed to be small, so that the mean square displacement is small by macroscopical standards but at the same time large compared to the interval between individual collisions in the case of gaseous diffusion, and for diffusion in liquids it may conveniently be identified with the mean life of a molecule in a given site. In both cases the velocity of the molecule at the end of the time interval is no longer correlated with that at the beginning of the interval, and the displacements in successive intervals are therefore independent of each other.

Concerning the local parameters it can be said that these are the least number of local macroscopic variables adequate for a complete specification of the macroscopic

state of a liquid mixture in the vicinity of a given point inside the fluid. Seeing that a drop of liquid mixture is but an ensemble of several kinds of molecules, besides their intrinsic properties such as the molecular masses (m_i), potential functions between pairs of molecules, one can further specify the average number of molecules per unit volume of each kind (n_i), their average momentum (\mathbf{u}), average kinetic energy (T) and the potential energy of each kind of molecules in the external force field (E_i).

For the purpose of calculating the coefficient of diffusion it will be seen in the following that only the first and second moments ($\overline{(\Delta \mathbf{x})_i}$, $\overline{(\Delta \mathbf{x})_i^2}$) of the transition probability ψ_i are required. In the case of self-diffusion it is well known (Einstein 1905) that the coefficient of self-diffusion D_{is} of the i th kind of molecule in an equilibrium mixture can be expressed as

$$D_{is} = \frac{1}{6} \frac{\overline{(\Delta \mathbf{x})_i^2}}{\Delta t}, \quad (1)$$

where
$$\overline{(\Delta \mathbf{x})_i^2} = \int \psi_i(\Delta \mathbf{x}, \Delta t) (\Delta \mathbf{x})^2 d(\Delta \mathbf{x}), \quad (2)$$

the integration being extended over all values of $(\Delta \mathbf{x})$. ψ_1 in the integrand depends only on $\Delta \mathbf{x}$, Δt and not on \mathbf{x} and t , as the λ_i 's are constant in a uniform fluid. For mutual diffusion in general one considers a 'normal' non-uniform fluid in which the λ_i 's vary smoothly in space. (Hereafter we limit ourselves to a binary mixture.)

Let $n_1^{(i)}(\mathbf{x}, t) d\mathbf{x}$ be the probability of finding the i th molecule of the first kind at t in \mathbf{x} , $d\mathbf{x}$. Then one can express the probability distribution of a given molecule in $t + \Delta t$ in terms of that at t and $\psi(\mathbf{x}, \Delta \mathbf{x}, t, \Delta t)$, viz.

$$n_1^{(i)}(\mathbf{x}, t + \Delta t) = \int n_1^{(i)}(\mathbf{x} - \Delta \mathbf{x}, t) \psi_1(\mathbf{x} - \Delta \mathbf{x}, \Delta \mathbf{x}, t, \Delta t) d(\Delta \mathbf{x}), \quad (3)$$

the integration being extended over the whole volume occupied by the fluid. Equation (3) can be expanded on both sides on account of our assumption made concerning Δt ,

$$\begin{aligned} \Delta t \frac{\partial}{\partial t} n_1^{(i)}(\mathbf{x}, t) + O(\Delta t)^2 \\ = - \frac{\partial}{\partial \mathbf{x}} \cdot [n_1^{(i)}(\mathbf{x}, t) \overline{(\Delta \mathbf{x})_1}] + \frac{1}{2} \frac{\partial}{\partial \mathbf{x}} \frac{\partial}{\partial \mathbf{x}} : [n_1^{(i)}(\mathbf{x}, t) \overline{(\Delta \mathbf{x} \Delta \mathbf{x})_1}] + O(\overline{\Delta \mathbf{x} \Delta \mathbf{x} \Delta \mathbf{x}}), \end{aligned} \quad (4)$$

where

$$\overline{(\Delta \mathbf{x})_1} = \int \psi_1(\mathbf{x}, \Delta \mathbf{x}, t, \Delta t) (\Delta \mathbf{x}) d(\Delta \mathbf{x}), \quad \overline{(\Delta \mathbf{x} \Delta \mathbf{x})_1} = \int \psi_1(\mathbf{x}, \Delta \mathbf{x}, t, \Delta t) (\Delta \mathbf{x} \Delta \mathbf{x}) d(\Delta \mathbf{x}). \quad (5)$$

Assuming that both $\overline{(\Delta \mathbf{x})_1}$ and $\overline{(\Delta \mathbf{x} \Delta \mathbf{x})_1}$ are of the order of Δt and neglecting terms of higher order than Δt , one obtains from equation (4)

$$\frac{\partial}{\partial t} n_1^{(i)}(\mathbf{x}, t) = - \frac{\partial}{\partial \mathbf{x}} \cdot [n_1^{(i)}(\mathbf{x}, t) \beta_1] + \frac{\partial}{\partial \mathbf{x}} \frac{\partial}{\partial \mathbf{x}} : [n_1^{(i)}(\mathbf{x}, t) \alpha_1], \quad (6)$$

where*

$$\beta_1 \Delta t = \overline{(\Delta \mathbf{x})_1}, \quad 2\alpha_1 \Delta t = \overline{(\Delta \mathbf{x} \Delta \mathbf{x})_1}. \quad (7)$$

* The α in clarendon type represents a tensor quantity.

Since ψ_1 and hence α_1 depends on the space gradients of λ_i at \mathbf{x}, t which are assumed to be small, we may expand α_1 with respect to the gradients, i.e.

$$\alpha_1 = \frac{1}{3}\alpha_1 1 + O\left(\frac{\partial \lambda_i}{\partial \mathbf{x}}\right), \quad \alpha_1 = \frac{1}{2}(\overline{\Delta \mathbf{x}})^2 / \Delta t, \quad (8)$$

where 1 stands for the unit tensor. As terms of higher order than the second in the gradients on the right-hand side of (6) will be neglected, we require for β_1 terms involving at most the first-order gradients and for α_1 only the first term in (8), viz. $\frac{1}{3}\alpha_1 1$. Equation (6) can then be written in the form

$$\frac{\partial}{\partial t} n_1^{(i)}(\mathbf{x}, t) = -\frac{\partial}{\partial \mathbf{x}} \cdot [n_1^{(i)}(\mathbf{x}, t) \beta_1] + \frac{1}{3} \frac{\partial}{\partial \mathbf{x}} \cdot \frac{\partial}{\partial \mathbf{x}} [n_1^{(i)}(\mathbf{x}, t) \alpha_1]. \quad (9)$$

It follows from the equation of continuity for $n_1^{(i)}(\mathbf{x}, t)$ that the local flow $n_1^{(i)}(\mathbf{x}, t) \mathbf{u}_1^{(i)}(\mathbf{x}, t)$ of the probability density $n_1^{(i)}(\mathbf{x}, t)$ is

$$n_1^{(i)} \mathbf{u}_1^{(i)} = n_1^{(i)}(\mathbf{x}, t) \beta_1 - \frac{1}{3} \frac{\partial}{\partial \mathbf{x}} (n_1^{(i)}(\mathbf{x}, t) \alpha_1). \quad (10)$$

Up to the present, only the probability of finding a particular molecule at a given position and time is considered. To relate it with the local partial density $n_1(\mathbf{x}, t)$ of molecules of the first kind, one can imagine that at some given instant when the mixture is already in its normal state, the λ_i 's are known throughout the fluid, and in particular that the positions of all the molecules are exactly known; the transition probability of each molecule is found by treating the rest as a continuous medium, with the λ_i 's known everywhere. At any later instant t , the probability of finding the volume element $\mathbf{x}, d\mathbf{x}$ occupied by any molecule of the first kind must be equal to the sum of the probabilities that it is occupied by each of the N_1 molecules of the first kind, i.e.

$$n_1(\mathbf{x}, t) = \sum_{i=1}^{N_1} n_1^{(i)}(\mathbf{x}, t). \quad (11)$$

Besides, α_1 and β_1 depend only on \mathbf{x}, t and are independent of which molecule of the first kind is chosen for consideration. The total local flow of molecules of the first kind is

$$n_1 \mathbf{u}_1 = \sum_{i=1}^{N_1} n_1^{(i)} \mathbf{u}_1^{(i)} = n_1 \beta_1 - \frac{1}{3} \frac{\partial}{\partial \mathbf{x}} (n_1 \alpha_1), \quad (12)$$

of which a part is due to the local mass velocity \mathbf{u} and the rest is the interdiffusion velocity \mathbf{u}'_1 . Hence

$$\mathbf{u}'_1 = (\beta_1 - \mathbf{u}) - \frac{1}{3n_1} \frac{\partial}{\partial \mathbf{x}} (n_1 \alpha_1) = \beta'_1 - \frac{1}{3} \frac{\partial \alpha_1}{\partial \mathbf{x}} - \frac{\alpha_1}{3n_1} \frac{\partial n_1}{\partial \mathbf{x}}, \quad (13)$$

where

$$\beta'_1 = \beta_1 - \mathbf{u}. \quad (14)$$

Proceeding in the same way, one obtains for molecules of the second kind

$$\mathbf{u}'_2 = \beta'_2 - \frac{1}{3} \frac{\partial \alpha_2}{\partial \mathbf{x}} - \frac{\alpha_2}{3n_2} \frac{\partial n_2}{\partial \mathbf{x}}, \quad (15)$$

where β'_2 , α_2 and n_2 have analogous meanings.

The mutual diffusion velocity $(\mathbf{u}'_1 - \mathbf{u}'_2)$ follows from (13) and (15):

$$\mathbf{u}'_1 - \mathbf{u}'_2 = (\beta'_1 - \beta'_2) - \frac{1}{3} \frac{\partial}{\partial \mathbf{x}} (\alpha_1 - \alpha_2) - \frac{1}{3} \left(\frac{\alpha_1}{n_1} \frac{\partial n_1}{\partial \mathbf{x}} - \frac{\alpha_2}{n_2} \frac{\partial n_2}{\partial \mathbf{x}} \right). \quad (16)$$

From the above equation one can regard the total mutual diffusion velocity $(\mathbf{u}'_1 - \mathbf{u}'_2)$ as consisting of three parts: the first, $(\beta'_1 - \beta'_2)$, is the difference of the rates of drift of the two kinds of molecules; the second, $-\frac{1}{3} \frac{\partial}{\partial \mathbf{x}} (\alpha_1 - \alpha_2)$, is the gradient of the difference of the rates of dispersion; and the third, $-\frac{1}{3} \left(\frac{\alpha_1}{n_1} \frac{\partial n_1}{\partial \mathbf{x}} - \frac{\alpha_2}{n_2} \frac{\partial n_2}{\partial \mathbf{x}} \right)$, is the contribution from self-diffusion in the presence of density gradients. As the evaluation of the β 's requires the real velocity distribution in a non-uniform fluid, the most that one can expect, without knowing the deviation from the Maxwellian distribution, is to determine the second and third parts in (16), and this is just what the old free-path theory was intended to do, though the persistence of motion has never been properly taken care of. This will be discussed in more detail in § 3.

In the following we shall give the formal expressions of the coefficients of ordinary diffusion D_0 and thermal diffusion D_T in terms of β'_1 , β'_2 , α_1 and α_2 . As β'_1 and β'_2 are linear in the gradients, so are $\partial \alpha_1 / \partial \mathbf{x}$ and $\partial \alpha_2 / \partial \mathbf{x}$, and the only components which are of interest in the present calculation are in the directions of $\partial n_1 / \partial \mathbf{x}$, $\partial n_2 / \partial \mathbf{x}$ and $\partial T / \partial \mathbf{x}$; one can write

$$\beta'_1 = \beta_1^{(1)} \frac{\partial n_1}{\partial \mathbf{x}} + \beta_1^{(2)} \frac{\partial n_2}{\partial \mathbf{x}} + \beta_1^{(3)} \frac{\partial T}{\partial \mathbf{x}}, \quad \frac{\partial \alpha_1}{\partial \mathbf{x}} = \frac{\partial \alpha_1}{\partial n_1} \frac{\partial n_1}{\partial \mathbf{x}} + \frac{\partial \alpha_1}{\partial n_2} \frac{\partial n_2}{\partial \mathbf{x}} + \frac{\partial \alpha_1}{\partial T} \frac{\partial T}{\partial \mathbf{x}}, \quad (17)$$

with similar expressions for β'_2 and $\partial \alpha_2 / \partial \mathbf{x}$. That the $\beta^{(i)}$'s ($i = 1, 2, 3$) are simply scalars is required by the fact that they can only be functions of n_1 , n_2 and T .

In order to select the part of the diffusion velocity due to the presence of a pressure gradient, one transforms the gradients $\partial n_1 / \partial \mathbf{x}$, $\partial n_2 / \partial \mathbf{x}$ and $\partial T / \partial \mathbf{x}$ to $\partial c_1 / \partial \mathbf{x}$, $\partial p / \partial \mathbf{x}$ and $\partial T / \partial \mathbf{x}$, where $c_1 = n_1 / n$ and $n = n_1 + n_2$, while $p = p_1 + p_2$ is the total pressure:

$$\frac{\partial c_1}{\partial \mathbf{x}} = \frac{n_1 n_2}{n^2} \left(\frac{1}{n_1} \frac{\partial n_1}{\partial \mathbf{x}} - \frac{1}{n_2} \frac{\partial n_2}{\partial \mathbf{x}} \right), \quad \frac{\partial p}{\partial \mathbf{x}} = \frac{\partial n_1}{\partial \mathbf{x}} \frac{\partial p}{\partial n_1} + \frac{\partial n_2}{\partial \mathbf{x}} \frac{\partial p}{\partial n_2} + \frac{\partial T}{\partial \mathbf{x}} \frac{\partial p}{\partial T}. \quad (18)$$

Hence one can express $\partial n_1 / \partial \mathbf{x}$ and $\partial n_2 / \partial \mathbf{x}$ in terms of $\partial c_1 / \partial \mathbf{x}$, $\partial p / \partial \mathbf{x}$ and $\partial T / \partial \mathbf{x}$:

$$\left. \begin{aligned} \frac{\partial n_1}{\partial \mathbf{x}} &= \left\{ n_2 \frac{\partial p}{\partial n_2} \frac{\partial c_1}{\partial \mathbf{x}} + n_1 \left(\frac{\partial p}{\partial \mathbf{x}} - \frac{\partial p}{\partial T} \frac{\partial T}{\partial \mathbf{x}} \right) \right\} / \mathcal{D} p, \\ \frac{\partial n_2}{\partial \mathbf{x}} &= \left\{ -n_2^2 \frac{\partial p}{\partial n_1} \frac{\partial c_1}{\partial \mathbf{x}} + n_2 \left(\frac{\partial p}{\partial \mathbf{x}} - \frac{\partial p}{\partial T} \frac{\partial T}{\partial \mathbf{x}} \right) \right\} / \mathcal{D} p, \end{aligned} \right\} \quad (19)$$

where
$$\mathcal{D} = n_1 \frac{\partial}{\partial n_1} + n_2 \frac{\partial}{\partial n_2}. \quad (20)$$

In the experimental observation of diffusion in liquids or gases the pressure is usually kept constant, thus one can take $\partial p / \partial \mathbf{x} = 0$ in (19). Substituting (19) in (17), one has

$$\left. \begin{aligned} \beta'_1 &= \frac{n_2^2 \left(\frac{\partial p}{\partial n_2} \beta_1^{(1)} - \frac{\partial p}{\partial n_1} \beta_1^{(2)} \right)}{\mathcal{D} p} \frac{\partial c_1}{\partial \mathbf{x}} + \left[\beta_1^{(3)} - \left(\frac{n_1 \beta_1^{(1)} + n_2 \beta_1^{(2)}}{\mathcal{D} p} \right) \frac{\partial p}{\partial T} \right] \frac{\partial T}{\partial \mathbf{x}}, \\ \frac{\partial \alpha_1}{\partial \mathbf{x}} &= \frac{n_2^2 \frac{\partial (\alpha_1, p)}{\partial (n_1, n_2)} \frac{\partial c_1}{\partial \mathbf{x}}}{\mathcal{D} p} + \left[\frac{\partial \alpha_1}{\partial T} - \frac{\mathcal{D} \alpha_1}{\mathcal{D} p} \frac{\partial p}{\partial T} \right] \frac{\partial T}{\partial \mathbf{x}}, \end{aligned} \right\} \quad (21)$$

with similar expressions for β'_2 and $\partial \alpha_2 / \partial \mathbf{x}$.

Equation (16), after inserting the formal expressions of $\beta'_1, \beta'_2, \partial\alpha_1/\partial\mathbf{x}$ and $\partial\alpha_2/\partial\mathbf{x}$, becomes

$$\begin{aligned} (\mathbf{u}'_1 - \mathbf{u}'_2) = & \frac{n^2}{\mathcal{D}p} \left\{ (\beta_1^{(1)} - \beta_2^{(1)}) \frac{\partial p}{\partial n_2} - (\beta_1^{(2)} - \beta_2^{(2)}) \frac{\partial p}{\partial n_1} - \frac{1}{3} \frac{\partial(\alpha_1 - \alpha_2, p)}{\partial(n_1, n_2)} - \frac{1}{3} \left(\frac{\alpha_1}{n_1} \frac{\partial p}{\partial n_2} + \frac{\alpha_2}{n_2} \frac{\partial p}{\partial n_1} \right) \right\} \frac{\partial c_1}{\partial \mathbf{x}} \\ & + \left\{ (\beta_1^{(3)} - \beta_2^{(3)}) - [n_1(\beta_1^{(1)} - \beta_2^{(1)}) + n_2(\beta_1^{(2)} - \beta_2^{(2)})] \frac{\partial p / \partial T}{\mathcal{D}p} \right. \\ & \left. + \frac{1}{3} \left[\frac{\partial p / \partial T}{\mathcal{D}p} (1 + \mathcal{D}) - \frac{\partial}{\partial T} \right] (\alpha_1 - \alpha_2) \right\} \frac{\partial T}{\partial \mathbf{x}}. \quad (22) \end{aligned}$$

Comparing with the usual definition of the coefficients of ordinary diffusion D_0 and thermal diffusion D_T ,

$$\mathbf{u}'_1 - \mathbf{u}'_2 = - \frac{n^2}{n_1 n_2} \left(D_0 \frac{\partial c_1}{\partial \mathbf{x}} - \frac{D_T}{T} \frac{\partial T}{\partial \mathbf{x}} \right), \quad (23)$$

one obtains finally

$$D_0 = \frac{n_1 n_2}{\mathcal{D}p} \left\{ \frac{1}{3} \frac{\partial(\alpha_1 - \alpha_2, p)}{\partial(n_1, n_2)} + \frac{1}{3} \left(\frac{\alpha_1}{n_1} \frac{\partial p}{\partial n_2} + \frac{\alpha_2}{n_2} \frac{\partial p}{\partial n_1} \right) - \left[(\beta_1^{(1)} - \beta_2^{(1)}) \frac{\partial p}{\partial n_2} - (\beta_2^{(2)} - \beta_1^{(2)}) \frac{\partial p}{\partial n_1} \right] \right\}, \quad (24)$$

$$\begin{aligned} D_T = \frac{n_1 n_2}{n^2} \left\{ \frac{1}{3} \left[\frac{T}{\mathcal{D}p} \frac{\partial}{\partial T} p (1 + \mathcal{D}) - T \frac{\partial}{\partial T} \right] (\alpha_1 - \alpha_2) \right. \\ \left. - [n_1(\beta_1^{(1)} - \beta_2^{(1)}) + n_2(\beta_1^{(2)} - \beta_2^{(2)})] \frac{T}{\mathcal{D}p} \frac{\partial}{\partial T} p + T(\beta_1^{(3)} - \beta_2^{(3)}) \right\}. \quad (25) \end{aligned}$$

In a dilute gas $p = nkT$, one has

$$D_0 = \frac{n_1 n_2}{n} \left\{ \frac{1}{3} \left(\frac{\partial}{\partial n_1} - \frac{\partial}{\partial n_2} \right) (\alpha_1 - \alpha_2) + \frac{1}{3} \left(\frac{\alpha_1}{n_1} + \frac{\alpha_2}{n_2} \right) - (\beta_1^{(1)} + \beta_2^{(2)}) + (\beta_2^{(1)} + \beta_1^{(2)}) \right\}, \quad (26)$$

$$D_T = \frac{n_1 n_2}{n^2} \left\{ \frac{1}{3} \left(1 + \mathcal{D} - T \frac{\partial}{\partial T} \right) (\alpha_1 - \alpha_2) - [n_1(\beta_1^{(1)} - \beta_2^{(1)}) + n_2(\beta_1^{(2)} - \beta_2^{(2)})] + T(\beta_1^{(3)} - \beta_2^{(3)}) \right\}. \quad (27)$$

3. STATISTICS OF THE MOTION OF GAS MOLECULES

In the present section it will be shown that one can actually follow the path of a single molecule statistically, assuming the real velocity distribution of the medium, as a function of \mathbf{x} and t , to be known. Advantage will be taken of the particularly simple feature in gases of low or moderate density, where the assumptions of molecular chaos and binary encounters are justified.

Strictly speaking one can define rigorously the free path only for rigid spheres. For any general force law the total cross-section calculated classically is always infinite owing to the weak interaction of the distant molecules, while in the quantum theory it is finite. For the present purpose we may cut off the interaction beyond a reasonable distance with negligible error, and it can then be shown that the final results α_1, α_2 , etc., are practically independent of the range of interaction which we have chosen at the beginning, so long as it is not too short (see § 4).

Complicated as the motion of a gas molecule is, it consists of only two kinds of events: it either travels freely or suffers a collision, with a certain probability of being deflected into a particular range of velocity after collision. Two probability functions which govern these two kinds of event will be defined first (in fact, it is necessary to determine only one probability, because one of them can be obtained from the other):

(1) $\mu(\mathbf{x}, \boldsymbol{\xi}^{(1)}, t) dt$ = probability that a molecule at \mathbf{x} of velocity $\boldsymbol{\xi}^{(1)}$ at time t will have suffered a collision in t, dt .

(2) $\chi(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \mathbf{x}, t) d\boldsymbol{\xi}^{(2)} dt$ = probability that a molecule at \mathbf{x} of velocity $\boldsymbol{\xi}^{(1)}$ at time t will have been deflected into $\boldsymbol{\xi}^{(2)}, d\boldsymbol{\xi}^{(2)}$ in t, dt .

It will be shown in appendix A that both μ and χ can be calculated for any force law with finite range.

It is obvious from the definitions of μ and χ that μ is obtainable from χ , thus

$$\mu(\boldsymbol{\xi}^{(1)}, \mathbf{x}, t) = \int \chi(\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \mathbf{x}, t) d\boldsymbol{\xi}^{(2)}. \quad (28)$$

Both μ and χ depend only on the instantaneous velocity $\boldsymbol{\xi}^{(1)}$ of the selected molecule at time t and not on its history; it does not matter how long the molecule has travelled with that velocity $\boldsymbol{\xi}^{(1)}$ before reaching \mathbf{x} at t . In a gas in equilibrium μ and χ are independent of \mathbf{x} and t ; μ is a function of $|\boldsymbol{\xi}^{(1)}|$ only and χ of $|\boldsymbol{\xi}^{(1)}|, |\boldsymbol{\xi}^{(2)}|$ and θ only, θ being the angle between $\boldsymbol{\xi}^{(1)}$ and $\boldsymbol{\xi}^{(2)}$.

In dealing with a series of successive flights, i.e. motions between encounters, we shall denote by $t_i, \mathbf{x}^{(i)}$ and $\boldsymbol{\xi}^{(i)}$ the time, position and velocity vectors at the beginning of the i th flight and by $\tau_{ij} = |t_i - t_j|$ the time interval between the i th and j th encounters, where the i th encounter occurs at the beginning of the i th flight.

Consider first the probability that a molecule at $\mathbf{x}^{(1)}$ with velocity $\boldsymbol{\xi}^{(1)}$ at time t_1 will travel freely for τ_{12} and suffer a collision in $\tau_{12}, d\tau_{12}$. Denoting this probability by $W(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\tau_{12}$, one has

$$W(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\tau_{12} = \left\{ 1 - \int_0^{\tau_{12}} W(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1, \tau) d\tau \right\} \mu(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(2)}, t_2) d\tau_{12}, \quad (29)$$

where the first factor on the right-hand side gives the probability that it will not collide before $t_2 = t_1 + \tau_{12}$ and the second factor the probability that it will collide in $\tau_{12}, d\tau_{12}$.

Dividing both sides of the above equation by $\mu(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(2)}, t_2)$ and differentiating with respect to τ_{12} , one obtains

$$\frac{d}{d\tau_{12}} \left(\frac{W(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12})}{\mu(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(2)}, t_2)} \right) = -W(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}).$$

Hence after integration one has

$$W(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) = \mu(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(2)}, t_2) \exp \left[- \int_0^{\tau_{12}} \mu(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)} + \boldsymbol{\xi}^{(1)}\tau, t_1 + \tau) d\tau \right], \quad (30)$$

using the initial condition

$$W(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1, 0) = \mu(\boldsymbol{\xi}^{(1)}, \mathbf{x}^{(1)}, t_1).$$

One notices here that the probability that a molecule of velocity $\xi^{(1)}$ at $\mathbf{x}^{(1)}$, t_1 will travel freely for an interval not less than τ_{12} without collision is according to (29)

$$\left\{ 1 - \int_0^{\tau_{12}} W(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau) d\tau \right\} = \exp \left[- \int_0^{\tau_{12}} \mu(\xi^{(1)}, \mathbf{x}^{(1)} + \xi^{(1)}\tau, t_1 + \tau) d\tau \right]. \quad (31)$$

Next the statistics of a complete flight is considered. (By a complete flight is meant the flight of a molecule between two successive collisions. The correct definition of the mean free path should be based on this concept. One sees clearly at this point the difference between Maxwell's free path which is in accord with the present definition and Tait's free path which is not.) Letting $W_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\xi^{(1)} d\mathbf{x}^{(1)} d\tau_{12}$ be the probability of finding a molecule at $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ at time t_1 with its velocity in $\xi^{(1)}$, $d\xi^{(1)}$ which travels freely for τ_{12} and suffers a collision in τ_{12} , $d\tau_{12}$, one has simply

$$W_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\xi^{(1)} d\mathbf{x}^{(1)} d\tau_{12} = f_1(\mathbf{x}^{(1)}, \xi^{(1)}, t_1) W(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\xi^{(1)} d\mathbf{x}^{(1)} d\tau_{12}, \quad (32)$$

where $f_1(\mathbf{x}^{(1)}, \xi^{(1)}, t_1)$ is the density in phase space of a single molecule at \mathbf{x} , t .

It should be noticed that among the molecules to be found in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ at t_1 with velocity in $\xi^{(1)}$, $d\xi^{(1)}$ some are just being deflected into $\xi^{(1)}$, $d\xi^{(1)}$, but by far the large majority of them have already acquired a velocity in $\xi^{(1)}$, $d\xi^{(1)}$ before t_1 , passing merely through $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ at t_1 . It would be incorrect to identify the average flight of the above set of molecules after t_1 with the free path of a single molecule, and this is just Tait's free path l_T in the direction of $\xi^{(1)}$

$$l_T = \frac{\iint W_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) \xi^{(1)} \tau_{12} \xi^{(1)3} d\xi^{(1)} d\tau_{12}}{\iint W_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) \xi^{(1)3} d\xi^{(1)} d\tau_{12}},$$

which for equilibrium state becomes

$$l_T = \overline{\left(\frac{\xi^{(1)}}{\mu(\xi^{(1)})} \right)},$$

where the bar denotes the average with respect to the Maxwellian velocity distribution.

In order to determine the real free path one has to select from the above set of molecules those which are just starting their new free paths in t_1 , dt_1 . This can be done by making use of the function χ . The probability of finding a molecule in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ at time t_1 with velocity in $\xi^{(1)}$, $d\xi^{(1)}$ which is deflected into $\xi^{(2)}$, $d\xi^{(2)}$ in t_1 , dt_1 is $d\mathbf{x}^{(1)} d\xi^{(1)} f_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) \chi(\xi^{(1)}, \xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(2)} dt_1$. Hence the probability that a molecule with any velocity in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ at time t_1 is deflected into $\xi^{(2)}$, $d\xi^{(2)}$ in t_1 , dt_1 is $d\mathbf{x}^{(1)} d\xi^{(2)} dt_1 \int f_1(\mathbf{x}^{(1)}, \xi^{(1)}, t_1) \chi(\xi^{(1)}, \xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(1)}$. It can be shown that in gases in equilibrium the above probability reduces to $f_1(\xi^{(2)}, \mathbf{x}^{(1)}, t_1) \mu(\xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(2)} d\mathbf{x}^{(1)} dt_1$, and in gases which are not in equilibrium it reduces to

$$\mu\left(\xi^{(2)}, \mathbf{x}^{(1)}, t_1 + \frac{\partial}{\partial t_1} + \xi^{(2)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}}\right) f_1(\xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(2)} d\mathbf{x}^{(1)} dt_1.$$

That this is correct in the equilibrium state can be inferred from the principle of detailed balancing. For, according to this principle, to every type of collision there exists the inverse kind occurring with equal frequency, exactly undoing the effect of the first. We therefore conclude that the probability of finding a molecule with any velocity at time t_1 in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ which is deflected into $\xi^{(2)}$, $d\xi^{(2)}$ in t_1 , dt_1 is equal to the probability of finding a molecule in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ at time t_1 with velocity in $\xi^{(2)}$, $d\xi^{(2)}$ which is deflected into a motion with any velocity in t_1 , dt_1 , i.e.

$$d\mathbf{x}^{(1)} d\xi^{(2)} dt_2 \int f_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) \chi(\xi^{(1)}, \xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(1)} \\ = f_1(\xi^{(2)}, \mathbf{x}^{(1)}, t_1) \mu(\xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(2)}, d\mathbf{x}^{(1)} dt_1. \quad (33)$$

For the non-equilibrium state the distribution function $f_1(\xi^{(2)}, \mathbf{x}^{(1)}, t_1)$ changes with time so that the set of molecules in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ with velocity in $\xi^{(2)}$, $d\xi^{(2)}$ increases in time dt_1 by the amount $\left(\frac{\partial}{\partial t_1} + \xi^{(2)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}}\right) f_1(\xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(2)} d\mathbf{x}^{(1)} dt_1$, which is just the extra term given above.

An alternative proof which may clarify the situation still further is as follows:

Let $P_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) dt_1 d\xi^{(1)} d\mathbf{x}^{(1)} d\tau_{12}$ be the probability of finding a molecule with any velocity in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ at time t_1 which is deflected into $\xi^{(2)}$, $d\xi^{(2)}$ in t_1 , dt_1 , travels freely for τ_{12} , and collides in τ_{12} , $d\tau_{12}$. Then it is seen that the number of the set of molecules $W_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\mathbf{x}^{(1)} d\xi^{(1)} d\tau_{12}$ must be equal to the integrated sum for all values of $t_0 < t_1$ of the number of molecules starting in $\mathbf{x}^{(0)}$, $d\mathbf{x}^{(0)}$ ($\mathbf{x}^{(0)} = \mathbf{x}^{(1)} - \xi^{(1)}\tau$) with velocity in $\xi^{(1)}$, $d\xi^{(1)}$, passing through $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ in t_1 , dt_1 , and arriving at $\mathbf{x}^{(2)}$, $d\mathbf{x}^{(2)}$ in t_2 , dt_2 . (The subscript 0 is used to indicate events prior to the instant t_1 .) Therefore

$$d\mathbf{x}^{(0)} d\xi^{(1)} d\tau_{02} \int P_1(\xi^{(1)}, \mathbf{x}^{(0)}, t_0, \tau_{02}) dt_0 = W_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\tau_{12} d\mathbf{x}^{(1)} d\xi^{(1)},$$

where the integration is to be carried out over all $t_0 < t_1$. If the time intervals are counted from t_2 , one can replace dt_0 , $d\tau_{02}$ and $d\tau_{12}$ by $d\tau_{02}$, dt_2 and dt_2 respectively. Thus

$$d\mathbf{x}^{(0)} d\xi^{(1)} dt_2 \int_{\tau_{12}}^{+\infty} P_1(\xi^{(1)}, \mathbf{x}^{(2)} - \xi^{(1)}\tau_{02}, t_2 - \tau_{02}, \tau_{02}) d\tau_{02} \\ = W(\xi^{(1)}, \mathbf{x}^{(2)} - \xi^{(1)}\tau_{12}, t_2 - \tau_{12}, \tau_{12}) d\xi^{(1)} d\mathbf{x}^{(1)} dt_2.$$

The volume elements $d\mathbf{x}^{(0)}$, $d\mathbf{x}^{(1)}$ can be chosen equal. After differentiating both sides with respect to τ_{12} , one obtains

$$-P_1(\xi^{(1)}, \mathbf{x}^{(2)} - \xi^{(1)}\tau_{12}, t_2 - \tau_{12}, \tau_{12}) = \frac{d}{d\tau_{12}} W_1(\xi^{(1)}, \mathbf{x}^{(2)} - \xi^{(1)}\tau_{12}, t_2 - \tau_{12}, \tau_{12}) \\ = \frac{d}{d\tau_{12}} \left\{ f_1(\xi^{(1)}, \mathbf{x}^{(2)} - \xi^{(1)}\tau_{12}, t_2 - \tau_{12}) \mu(\xi^{(1)}, \mathbf{x}^{(2)}, t_2) \exp \left[- \int_0^{\tau_{12}} \mu(\xi^{(1)}, \mathbf{x}^{(2)} - \xi^{(1)}\tau, t_2 - \tau) d\tau \right] \right\}.$$

$$\text{Hence} \quad P_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) = \left\{ \left(\mu(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) + \frac{\partial}{\partial t_1} + \xi^{(1)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}} \right) f_1(\mathbf{x}^{(1)}, \xi^{(1)}, t_1) \right\} \\ \times \mu(\mathbf{x}^{(2)}, \xi^{(1)}, t_2) \exp \left[- \int_0^{\tau_{12}} \mu(\xi^{(1)}, \mathbf{x}^{(1)} + \xi^{(1)}\tau, t + \tau) d\tau \right]. \quad (34)$$

By definition of χ and with the help of (34), one has

$$\begin{aligned} d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \int f_1(\xi^{(0)}, \mathbf{x}^{(1)}, t_1) \chi(\xi^{(0)}, \xi^{(1)}, \mathbf{x}^{(1)}, t_1) d\xi^{(0)} &= d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \int P_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) d\tau_{12} \\ &= d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \left(\mu(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) + \frac{\partial}{\partial t_1} + \xi^{(1)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}} \right) f_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1), \end{aligned} \quad (35)$$

which is the relation to be proved.

It should be added that equation (35) is merely another form of Maxwell and Boltzmann's equation. For using (28) one has

$$\begin{aligned} &\left(\frac{\partial}{\partial t_1} + \xi^{(1)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}} \right) f_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) \\ &= \int \{ f_1(\xi^{(0)}, \mathbf{x}^{(1)}, t_1) \chi(\xi^{(0)}, \xi^{(1)}, \mathbf{x}^{(1)}, t_1) - f_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) \chi(\xi^{(1)}, \xi^{(0)}, \mathbf{x}^{(1)}, t_1) \} d\xi^{(0)}. \end{aligned} \quad (36)$$

Having obtained $P_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12})$ in (34), one can define the real free path l_M for the direction $\xi^{(1)}$ as

$$l_M = \frac{\iint P_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) \xi^{(1)} \tau_{12} \xi^{(1)2} d\xi^{(1)} d\tau_{12}}{\iint P_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}) \xi^{(1)2} d\xi^{(1)} d\tau_{12}}.$$

For the equilibrium state it becomes the usual Maxwellian free path l_M

$$l_M = \frac{\overline{\xi^{(1)}}}{\mu(\overline{\xi^{(1)}})},$$

where the bars denote the average with respect to the Maxwellian velocity distribution.

The extension from a single flight to a series of successive flights is immediate. For two steps one defines $P_2(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}; \xi^{(2)}, \mathbf{x}^{(2)}, t_2, \tau_{23}) d\xi^{(1)} d\mathbf{x}^{(1)} dt_1 d\xi^{(2)} dt_2 d\tau_{23}$ as the probability of finding at t_1 in $\mathbf{x}^{(1)}$ with any velocity which is deflected into $\xi^{(1)}$, $d\xi^{(1)}$ in t_1 , dt_1 travels freely for an interval τ_{12} , is deflected again into $\xi^{(2)}$, $d\xi^{(2)}$ in t_2 , dt_2 , travels freely for another interval τ_{23} and finally collides in τ_{23} , $d\tau_{23}$. Then one can write

$$\begin{aligned} &P_2(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}; \xi^{(2)}, \mathbf{x}^{(2)}, t_2, \tau_{23}) d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 d\xi^{(2)} dt_2 d\tau_{23} \\ &= \left\{ \left(\mu(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) + \frac{\partial}{\partial t_1} + \xi^{(1)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}} \right) f_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \right\} \\ &\quad \times \left\{ \exp \left[- \int_0^{\tau_{12}} \mu(\xi^{(1)}, \mathbf{x}^{(1)} + \xi^{(1)} \tau, t_1 + \tau) d\tau \right] \chi(\xi^{(1)}, \xi^{(2)}, \mathbf{x}^{(1)}, t_1) d\xi^{(2)} dt_2 \right\} \\ &\quad \times \left\{ \exp \left[- \int_0^{\tau_{23}} \mu(\xi^{(2)}, \mathbf{x}^{(2)} + \xi^{(2)} \tau, t_2 + \tau) d\tau \right] \mu(\xi^{(2)}, \mathbf{x}^{(2)}, t_2) d\tau_{23} \right\}, \end{aligned} \quad (37)$$

where the first factor gives the probability of finding a molecule in $\mathbf{x}^{(1)}$, $d\mathbf{x}^{(1)}$ with velocity in $\xi^{(1)}$, $d\xi^{(1)}$, starting its free path in t_1 , dt_1 , the second factor gives the probability that it will travel freely for τ_{12} , be deflected into $\xi^{(2)}$, $d\xi^{(2)}$ in t_2 , dt_2 , and the last factor

gives the probability that it will travel again for τ_{23} and suffer a collision in τ_{23} , $d\tau_{23}$.

By making use of the properties of χ (28), (35) it can be shown that

$$\int \int P_2(12) d\xi^{(2)} d\tau_{23} = P_1(1), \quad \int \int P_2(12) d\xi^{(1)} dt_1 = P_1(2),$$

where $P_1(2)$ and $P_2(12)$ are used to denote

$$P_1(\xi^{(1)}, \mathbf{x}^{(2)}, t_2, \tau_{23}) \quad \text{and} \quad P_2(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}; \xi^{(2)}, \mathbf{x}^{(2)}, t_2, \tau_{23})$$

respectively.

Similarly, the probability function of N successive flights can be written down as

$$\begin{aligned} P_N(\xi^{(1)}, \mathbf{x}^{(1)}, t_1, \tau_{12}, \dots, \xi^{(N)}, \mathbf{x}^{(N)}, t_N, \tau_{N,N+1}) d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \dots d\xi^{(N)} dt_N d\tau_{N,N+1} \\ = P_N(1, 2, \dots, N) d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \dots d\xi^{(N)} dt_N d\tau_{N,N+1} \\ = \left\{ \left(\mu(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) + \frac{\partial}{\partial t_1} + \xi^{(1)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}} \right) f_1(\xi^{(1)}, \mathbf{x}^{(1)}, t_1) d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \right\} \\ \times \prod_{i=1}^{N-1} \left\{ \exp \left[- \int_0^{\tau_{i,i+1}} \mu(\xi^{(i)}, \mathbf{x}^{(i)} + \xi^{(i)} \tau, t_i + \tau) d\tau \right] \chi(\xi^{(i)}, \xi^{(i+1)}, \mathbf{x}^{(i+1)}, t_{i+1}) d\xi^{(i+1)} dt_{i+1} \right\} \\ \times \frac{d}{d\tau_{N,N+1}} \left\{ \exp \left[- \int_0^{\tau_{N,N+1}} \mu(\xi^{(N)}, \mathbf{x}^{(N)} + \xi^{(N)} \tau, t_N + \tau) d\tau \right] \right\} d\tau_{N,N+1}. \end{aligned} \quad (38)$$

As expected the following relations are automatically satisfied by $P_N(1, 2, \dots, N)$:

$$\begin{aligned} \int \int P_N(1, 2, \dots, N) d\xi^{(N)} d\tau_{N,N+1} &= P_{N-1}(1, 2, \dots, N-1), \\ \int \int P_N(1, 2, \dots, N) d\xi^{(1)} dt_1 &= P_{N-1}(2, 3, \dots, N), \end{aligned}$$

or more generally

$$\begin{aligned} \int \dots \int P_N(1, 2, \dots, N) dt_1 d\xi^{(1)} \dots dt_r d\xi^{(r)} d\xi^{(s)} d\tau_{s,s+1} \dots d\xi^{(N)} d\tau_{N,N+1} \\ = P_{s-r-1}(r+1, \dots, s-1) (s-r \geq 2). \end{aligned}$$

For the equilibrium state, P_N reduces to

$$\begin{aligned} P_N(1, 2, \dots, N) d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \dots d\xi^{(N)} dt_N d\tau_{N,N+1} \\ = \{ \mu(\xi^{(1)}) f(\xi^{(1)}) d\mathbf{x}^{(1)} d\xi^{(1)} dt_1 \} \prod_{i=1}^{N-1} \{ \exp [- \mu(\xi^{(i)}) \tau_{i,i+1}] \chi(\xi^{(i)}, \xi^{(i+1)}) d\xi^{(i+1)} dt_{i+1} \} \\ \times \frac{d}{d\tau_{N,N+1}} \{ \exp [- \mu(\xi^{(N)}) \tau_{N,N+1}] \} d\tau_{N,N+1}. \end{aligned} \quad (39)$$

4. SELF-DIFFUSION IN GASES

Having found $P_N(1, 2, \dots, N)$ in (39) for a gas in equilibrium, one can calculate the mean time of flight Δt and mean square displacement $(\overline{\Delta \mathbf{x}})^2$ for any number N of successive flights. To conform with the condition imposed on Δt in the last section, we choose a large number of flights so that $(\overline{\Delta \mathbf{x}})^2/\Delta t$ approaches a plateau value. It can be shown that

$$\Delta t = \left\langle \sum_{i=1}^N \tau_{i,i+1} \right\rangle = N \langle \tau_{12} \rangle = \frac{N}{\mu(1)}, \quad (40)$$

where $\overline{\mu(1)} = \int \frac{f(1)}{n} \mu(1) d\xi^{(1)}$, $\mu(1) = \mu(\xi^{(1)})$, $f(1) = f(\xi^{(1)})$,

and n is the number density; the angular bracket $\langle \rangle$ is used to denote the average with respect to P_N .

To prove (40), one has to calculate $\langle \tau_{i,i+1} \rangle$, using (39) and (33),

$$\begin{aligned} \langle \tau_{i,i+1} \rangle &= \frac{\int \dots \int P_i(1, 2, \dots, i) \tau_{i,i+1} d\xi^{(1)} d\tau_{12} \dots d\xi^{(i)} d\tau_{i,i+1}}{\int \dots \int P_i(1, 2, \dots, i) d\xi^{(1)} d\tau_{12} \dots d\xi^{(i)} d\tau_{i,i+1}} \\ &= \frac{1}{n \overline{\mu(1)}} \int \dots \int f(1) \frac{\chi(12) \dots \chi(i-1, i)}{\mu(2) \dots \mu(i)} d\xi^{(1)} \dots d\xi^{(i)} = \frac{1}{\mu(1)}, \end{aligned}$$

where

$$\chi(12) = \chi(\xi^{(1)}, \xi^{(2)}).$$

Let $\mathbf{l}^{(i)} = \xi^{(i)} \tau_{i,i+1}$, then the mean displacement $\overline{(\Delta \mathbf{x})} = \left\langle \sum_{i=1}^N \mathbf{l}^{(i)} \right\rangle$ obviously vanishes by symmetry. The mean-square displacement $\overline{(\Delta \mathbf{x})^2} = \left\langle \left(\sum_{i=1}^N \mathbf{l}^{(i)} \right)^2 \right\rangle$ will next be shown to be

$$\begin{aligned} \overline{(\Delta \mathbf{x})^2} &= \left\langle \left(\sum_{i=1}^N \mathbf{l}^{(i)} \right)^2 \right\rangle = \frac{2N}{\mu(1)} \left\{ \int \frac{f(1)}{n} \frac{\xi^{(1)2}}{\mu(1)} d\xi^{(1)} + \left(\frac{N-1}{N} \right) \iint \frac{f(1)}{n} \frac{\chi(12)}{\mu(1)\mu(2)} \xi^{(1)} \cdot \xi^{(2)} d\xi^{(1)} d\xi^{(2)} \right. \\ &\quad \left. + \left(\frac{N-2}{N} \right) \iiint \frac{f(1)}{n} \frac{\chi(12)\chi(23)}{\mu(1)\mu(2)\mu(3)} \xi^{(1)} \cdot \xi^{(3)} d\xi^{(1)} d\xi^{(2)} d\xi^{(3)} + \dots \right\}. \quad (41) \end{aligned}$$

We have identified $\left\langle \sum_{i=1}^N \mathbf{l}^{(i)} \right\rangle$ and $\left\langle \left(\sum_{i=1}^N \mathbf{l}^{(i)} \right)^2 \right\rangle$ obtained from P_N with $\overline{(\Delta \mathbf{x})}$ and $\overline{(\Delta \mathbf{x})^2}$ from ψ without comment. In fact they are really slightly different; while in the definition of P_N the first flight definitely starts just after the last collision, the function ψ only specifies the probability of a displacement from a given initial position which may be any intermediate stage in a free flight. However, such difference becomes negligible when one considers what happens in a large number of free flights.

The meaning of the successive terms, apart from the constant coefficients $\left(\frac{N-i}{N} \right)$ ($i = 1, 2, \dots$), inside the bracket, in (41) is this: the first term inside the bracket is the main contribution from each individual step, the second term results from the correlation between two successive steps, the third term from the correlation between two steps separated by one flight, etc. Physically it is clear that such correlation decreases rapidly for each step, and the above series is therefore a rapidly convergent one.

To prove (41), one writes

$$\left\langle \left(\sum_{i=1}^N \mathbf{l}^{(i)} \right)^2 \right\rangle = \sum_{i=1}^N \langle (\mathbf{l}^{(i)})^2 \rangle + 2 \sum_{i=1}^{N-1} \langle (\mathbf{l}^{(i)} \cdot \mathbf{l}^{(i+1)}) \rangle + 2 \sum_{i=1}^{N-2} \langle (\mathbf{l}^{(i)} \cdot \mathbf{l}^{(i+2)}) \rangle + \dots \quad (42)$$

So, using (39) and (33), one has

$$\begin{aligned}
 \langle I^{(i)} \rangle &= \frac{\int \dots \int P_i(1, 2, \dots, i) (\xi^{(i)} \tau_{i, i+1})^2 d\xi^{(1)} d\tau_{12} \dots d\xi^{(i)} d\tau_{i, i+1}}{\int \dots \int P_i(1, 2, \dots, i) d\xi^{(1)} d\tau_{12} \dots d\xi^{(i)} d\tau_{i, i+1}} \\
 &= \frac{2}{\mu(1)} \int \frac{f(1)}{n} \frac{\xi^{(1)2}}{\mu(1)} d\xi^{(1)}, \\
 \langle I^{(i)} \cdot I^{(i+k)} \rangle &= \frac{\int \dots \int P_{i+k}(1, 2, \dots, i+k) \xi^{(i)} \cdot \xi^{(i+k)} \tau_{i, i+i} \tau_{i+k, i+k+1} d\xi^{(1)} d\tau_{12} \dots d\xi^{(i+k)} d\tau_{i+k, i+k+1}}{\int \dots \int P_{i+k}(1, 2, \dots, i+k) d\xi^{(1)} d\tau_{12} \dots d\xi^{(i+k)} d\tau_{i+k, i+k+1}} \\
 &= \frac{1}{n\mu(1)} \int \dots \int f(i) \frac{\chi(i, i+1) \dots \chi(i+k-1, i+k)}{\mu(i) \dots \mu(i+k)} \xi^{(i)} \cdot \xi^{(i+k)} d\xi^{(i)} \dots d\xi^{(i+k)} \\
 &= \frac{1}{\mu(1)} \int \dots \int \frac{f(1)}{n} \frac{\chi(12) \dots \chi(k, k+1)}{\mu(1) \dots \mu(k+1)} \xi^{(1)} \cdot \xi^{(k+1)} d\xi^{(1)} \dots d\xi^{(k+1)}.
 \end{aligned} \tag{43}$$

The above expression will give for $k = 1, 2, \dots$ the second, third, etc., terms in (42). Equation (41) follows therefore from (42) and (43). From (40) and (41), it is seen that for large N the above ratio of the mean square displacement to the time of travel approaches a plateau value, and the coefficient of self-diffusion is therefore

$$\begin{aligned}
 D_s = \frac{1}{6} \frac{(\Delta x)^2}{\Delta t} &= \frac{1}{3} \left\{ \int \frac{f(1)}{n} \frac{\xi^{(1)2}}{\mu(1)} d\xi^{(1)} + \int \frac{f(1)}{n} \frac{\chi(12)}{\mu(1)\mu(2)} \xi^{(1)} \cdot \xi^{(2)} d\xi^{(1)} d\xi^{(2)} \right. \\
 &\quad \left. + \int \int \frac{f(1)}{n} \frac{\chi(12)\chi(23)}{\mu(1)\mu(2)\mu(3)} \xi^{(1)} \cdot \xi^{(3)} d\xi^{(1)} d\xi^{(2)} d\xi^{(3)} + \dots \right\}. \tag{44}
 \end{aligned}$$

The above calculation refers to pure self-diffusion in an equilibrium gas consisting of only one kind of molecules. The extension to an equilibrium gas mixture is immediate. Consider a binary gas mixture in equilibrium. Let n_1 and n_2 be the number densities of the two kinds of molecules, $f_1(i)$ ($= f_1(\xi_1^{(i)})$) and $f_2(i)$ ($= f_2(\xi_2^{(i)})$) the densities in their molecular phase spaces, and $\mu_1(i)$ ($= \mu_1(\xi_1^{(i)})$) and $\mu_2(i)$ ($= \mu_2(\xi_2^{(i)})$) the total collision frequencies of a molecule of the first kind with velocity $\xi^{(1)}$ and of a molecule of the second kind with velocity $\xi^{(2)}$ respectively. One has

$$\mu_1(i) = \mu_{11}(i) + \mu_{12}(i), \tag{45}$$

where $\mu_{11}(i)$ represents collision between molecules of the first kind and $\mu_{12}(i)$ between molecules of the first kind and second kind. Using the same type of subscript, we have for the deflexion probability χ

$$\chi_1(i, i+1) = \chi_{11}(i, i+1) + \chi_{12}(i, i+1). \tag{46}$$

We can then write, for the coefficient of self-diffusion of the first kind of molecules

$$\begin{aligned}
 D_{1s} &= \frac{1}{3} \left\{ \int \frac{f_1(1)}{n_1} \frac{\xi_1^{(1)2}}{\mu_1(1)} d\xi_1^{(1)} + \int \frac{f_1(1)}{n_1} \frac{\chi_1(12)}{\mu_1(1)\mu_1(2)} \xi_1^{(1)} \cdot \xi_1^{(2)} d\xi_1^{(1)} d\xi_1^{(2)} \right. \\
 &\quad \left. + \int \int \frac{f_1(1)}{n_1} \frac{\chi_1(12)\chi_1(23)}{\mu_1(1)\mu_1(2)\mu_1(3)} \xi_1^{(1)} \cdot \xi_1^{(3)} d\xi_1^{(1)} d\xi_1^{(2)} d\xi_1^{(3)} + \dots \right\}. \tag{47}
 \end{aligned}$$

It is interesting to note that while in the usual gas theory the coefficient of self-diffusion is obtained from the coefficient of ordinary diffusion in a gas mixture by making the two different types of molecules identical, a rather indirect way to achieve the purpose, the present theory enables it to be calculated directly from a Maxwellian distribution. In particular, while in the usual gas theory to calculate the coefficient of self-diffusion in a binary mixture one has to consider a ternary mixture which is much more complicated, the present theory is able to deal with a gas mixture in quite a similar way, as for a simple gas.

In order to see the meaning of the various terms in (44) more clearly we shall rewrite it as follows:

$$\begin{aligned} \overline{(\Delta \mathbf{x})^2} &= \left\langle \left(\sum_{i=1}^N \mathbf{l}^{(i)} \right)^2 \right\rangle = \sum_{i=1}^N \langle l^{(i)2} \rangle + \sum_{i=1}^{N-1} \langle 2\mathbf{l}^{(i)} \cdot \mathbf{l}^{(i+1)} \rangle + \sum_{i=1}^{N-2} \langle 2\mathbf{l}^{(i)} \cdot \mathbf{l}^{(i+2)} \rangle + \dots \\ &= N \langle l^{(1)2} \rangle + (N-1) \langle 2\mathbf{l}^{(1)} \cdot \mathbf{l}^{(2)} \rangle + (N-2) \langle 2\mathbf{l}^{(1)} \cdot \mathbf{l}^{(3)} \rangle + \dots, \\ D_s &= \frac{1}{6} \frac{\overline{(\Delta \mathbf{x})^2}}{\Delta t} = \frac{\langle l^{(1)2} \rangle}{6 \langle \tau_{12} \rangle} \left\{ 1 + \frac{\langle 2\mathbf{l}^{(1)} \cdot \mathbf{l}^{(2)} \rangle}{\langle l^{(1)2} \rangle} + \frac{\langle 2\mathbf{l}^{(1)} \cdot \mathbf{l}^{(3)} \rangle}{\langle l^{(1)2} \rangle} + \dots \right\}. \end{aligned} \quad (48)$$

Equation (48) shows clearly that the various terms of the series merely take into account the effect of the presistence of motion on the rate of diffusion in a rigorous way.

To see that the value of D_s is not appreciably changed by choosing different ranges of interaction for any general force law between molecules when the chosen range is not too small, we rewrite (48) in the form

$$D_s = \frac{1}{6} \frac{\langle l^{(1)2} \rangle^{\frac{1}{2}}}{\langle \tau_{12} \rangle} \left\{ \langle l^{(1)2} \rangle^{\frac{1}{2}} + \frac{\langle 2\mathbf{l}^{(1)} \cdot \mathbf{l}^{(2)} \rangle}{\langle l^{(1)2} \rangle^{\frac{1}{2}}} + \frac{\langle 2\mathbf{l}^{(1)} \cdot \mathbf{l}^{(3)} \rangle}{\langle l^{(1)2} \rangle^{\frac{1}{2}}} + \dots \right\}. \quad (49)$$

Increasing the range of interaction will tend to reduce the length of free path and mean life of free flight but leave their ratio, the average speed, almost unchanged.

This is approximately the factor $\frac{\langle l^{(1)2} \rangle^{\frac{1}{2}}}{\langle \tau_{12} \rangle}$ in (49): As to the series inside the bracket in (49), its sum is almost constant for a sufficiently long range of interaction, though the rate of convergence does depend on this choice of range of interaction. The exact nature of the convergence cannot be discussed until μ and χ have been calculated for some particular force law, but qualitatively one can see that as the range of interaction is increased, the first term tends to decrease, but then a small angular deflexion between successive flights becomes more probable, hence the second term increases, etc. A convenient limit for the range of interaction can be set up as follows: it must be larger than two to three times the molecular diameter found, for example, from measurements of the second virial coefficient in gases, but it must not extend up to a distance comparable with the average distance between the molecules corresponding to the density of the gas considered. Formula (44) is valid so long as these two limits do not come close to each other, which is always so in gases. Choosing a high limit to the range of interaction will give very slightly more accurate results, but this is entirely offset by the slowness of the convergence of the series. Numerical

calculations of the coefficient of self-diffusion for rigid sphere gas molecules are shown in appendix C.

5. ORDINARY AND THERMAL DIFFUSION IN GAS MIXTURES

It has been found possible in § 2 to express the mutual diffusion velocity in non-uniform fluids in terms of the first and second moments of certain transition probabilities. It has also been shown in § 3 that in non-equilibrium gases one can follow the future course of a chosen molecule provided the velocity distribution of the medium is known. The essential point is that the statistics of the motion of gas molecules is based on the knowledge of the velocity distribution throughout the medium. The mutual diffusion velocity according to (22) with β 's and α 's calculated from P_N (38) should be, and as shown in appendix B, is indeed equal to that calculated directly from the local velocity distribution, i.e.

$$\mathbf{u}'_1 - \mathbf{u}'_2 = \frac{1}{n_1} \int f'_1(\xi_1) \xi_1 d\xi_1 - \frac{1}{n_2} \int f'_2(\xi_2) \xi_2 d\xi_2, \quad (50)$$

where $f'_1(\xi_1)$ and $f'_2(\xi_2)$ are the deviations from the Maxwellian velocity distribution due to the presence of the first-order gradients of densities, temperature, etc., of molecules of the first and second kind respectively.

Nevertheless, the present calculation shows the intrinsically different natures of self-diffusion on the one hand and mutual diffusion on the other; the former is simply a kind of Brownian motion, while the latter is entirely caused by the non-uniformity of the physical parameters. Though in non-uniform gas mixtures self-diffusion still goes on, it has no effect on the diffusion velocity which can be pictured as set up by forces arising from the local gradients of densities, temperature, etc. In order to obtain an exact value of the mutual diffusion velocity it is indispensable to consider the local deviation from Maxwellian velocity distribution consistent with the local gradients of the physical parameters.

When the β 's are neglected in (22), the remaining part is seen to correspond to the free-path theory. Though the present theory which takes account of the persistence of successive flights rigorously is much more refined than the old free-path theory, they are both based on the same assumption that molecules which collide during each element of time have a distribution after collision of the Maxwellian type. Mutual diffusion, though entirely different from self-diffusion in nature, can thus be approximately expressed in terms of self-diffusion. Equation (22) becomes when the β 's are neglected

$$\mathbf{u}'_1 - \mathbf{u}'_2 = -\frac{1}{3} \left\{ \left(\frac{\partial n_1}{\partial \mathbf{x}} \frac{\partial}{\partial n_1} + \frac{\partial n_2}{\partial \mathbf{x}} \frac{\partial}{\partial n_2} + \frac{\partial T}{\partial \mathbf{x}} \frac{\partial}{\partial T} \right) (\alpha_1 - \alpha_2) + \left(\frac{\alpha_1}{n_1} \frac{\partial n_1}{\partial \mathbf{x}} - \frac{\alpha_2}{n_2} \frac{\partial n_2}{\partial \mathbf{x}} \right) \right\}. \quad (51)$$

With the aid of this equation both ordinary and thermal diffusion receive simple explanations. The α_1 and α_2 appearing in (51) measure the tendency to diffuse for the two kinds of molecules. Mutual diffusion can be seen from (51) to arise in two ways. First, the number of molecules diffusing in one direction is not equal to that in the opposite direction, and secondly, the tendency for diffusion varies in different ways

with n_1, n_2 and T for the two kinds of molecules. Equations (26) and (27) become, when the β 's are neglected,

$$\left. \begin{aligned} D_0 &= \frac{n_1 n_2}{3n} \left\{ \left(\frac{\partial}{\partial n_1} - \frac{\partial}{\partial n_2} \right) (\alpha_1 - \alpha_2) + \left(\frac{\alpha_1}{n_1} + \frac{\alpha_2}{n_2} \right) \right\}, \\ D_T &= \frac{n_1 n_2}{3n^2} \left\{ 1 + n_1 \frac{\partial}{\partial n_1} + n_2 \frac{\partial}{\partial n_2} - T \frac{\partial}{\partial T} \right\} (\alpha_1 - \alpha_2). \end{aligned} \right\} \quad (52)$$

Thus D_0 consists of contributions from both of the two processes described above under constant temperature and pressure, while D_T arises only from the second process under constant pressure. It may be mentioned here that the old free-path theory does not explain thermal diffusion because there only the first effect is taken into account. Meyer's diffusion coefficient (cf. Jeans 1925) corresponds to

$$D_0 = \frac{n_1 n_2}{3n} \left(\frac{\alpha_1}{n_1} + \frac{\alpha_2}{n_2} \right). \quad (53)$$

Fürth's explanation of thermal diffusion (Fürth 1942) corresponds to the second equation in (52).

It should be mentioned that the neglect of the β 's is no more justified than the whole procedure of the old free-path theory. For gas mixtures in which one component is comparatively rare, equation (52) would give good results. This suggests that a corresponding approximate theory exists in the liquid region which can be obtained without any knowledge of the deviation from Maxwellian velocity distribution, provided the coefficient of self-diffusion in an equilibrium liquid mixture is known. This is now under investigation.

The author wishes to express his thanks to Professor M. Born for his constant interest and encouragement, and to Dr H. S. Green for many helpful suggestions and discussions and for reading the manuscript.

APPENDIX A. CALCULATION OF μ AND χ FOR SPECIAL MODELS

As the collision frequency μ and the deflexion probability χ are independent of the frame of reference chosen, we shall assume a frame of reference moving with the local mass motion velocity \mathbf{u} . From (45) and (46), it is seen that each of $\mu_1(1)$ and $\chi_1(12)$ consists of two parts. It is obviously sufficient to calculate $\mu_{12}(1)$ and $\chi_{12}(12)$. We shall denote by $\mathbf{v}_1^{(1)}$ and $\mathbf{v}_1^{(2)}$ the velocities of a molecule of the first kind before and after a collision respectively and $\mathbf{v}_2^{(1)}$ and $\mathbf{v}_2^{(2)}$ those of a molecule of the second kind. Further, each of $\mu_{12}(1)$ and $\chi_{12}(12)$ can be expanded

$$\mu_{12}(1) = \mu_{12}^0(1) + \mu'_{12}(1) + \dots, \quad \chi_{12}(12) = \chi_{12}^0(12) + \chi'_{12}(12) + \dots,$$

where the first term arises from collisions with those of the second kind of molecules in the medium with the Maxwellian distribution, and the second term from those deviating from the Maxwellian distribution due to the presence of the gradients $\partial \lambda_i / \partial \mathbf{x}$. As $\mu'_{12}(1)$ and $\chi'_{12}(12)$ are not required in the present paper, only $\mu_{12}^0(1)$ and $\chi_{12}^0(12)$ will now be calculated.

Rigid sphere model

The collision frequency of rigid spherical molecules calculated on the assumption of binary encounter and molecular chaos is given in most text-books on the kinetic theory of gases (cf. Chapman & Cowling 1939). The result is

$$\mu_{12}^0(1) = \left(\frac{2\pi kT}{m_2} \right)^{\frac{1}{2}} \sigma_{12}^2 n_2 E(y_2), \quad (\text{A } 1)$$

where

$$\sigma_{12} = \frac{1}{2}(\sigma_1 + \sigma_2), \quad y_2 = \sqrt{\left(\frac{m^2}{2kT} \right)} v_1^{(1)}, \quad E(y_2) = \left\{ e^{-y_2^2} + \left(2y_2 + \frac{1}{y_2} \right) \int_0^{y_2} e^{-x^2} dx \right\}, \quad (\text{A } 2)$$

σ_1 and σ_2 are the molecular diameters of the two kinds of molecules and m_1 and m_2 the molecular masses.

To calculate $\chi_{12}^0(12)$, we proceed at first in a similar manner as in the calculation of $\mu_{12}^0(1)$, that is, by counting the probability of the occurrence of a specific type of collision but then integrating over all possible types of collision for which the chosen molecule is deflected into a specific velocity range $\mathbf{v}_1^{(2)}$, $d\mathbf{v}_1^{(2)}$ after collision. Thus one obtains

$$\chi_{12}^0(12) dt d\mathbf{v}_1^{(2)} = \int_R f_2^0(\mathbf{v}_2^{(1)}) g d\epsilon b db dt d\mathbf{v}_2^{(1)}, \quad (\text{A } 3)$$

where $f_2^0(\mathbf{v}_2^{(1)})$ is the Maxwellian velocity distribution of the second type of molecules normalized to its number density, \mathbf{g} ($= \mathbf{v}_2^{(1)} - \mathbf{v}_1^{(1)}$) is the velocity of a molecule in the medium relative to the chosen molecule, $b db d\epsilon$ is an elementary cross-section or 'target area' in polar co-ordinates on a plane passing through the chosen molecule and perpendicular to \mathbf{g} , b being the impact parameter and ϵ the polar angle. The integration is to be carried out over a region R defined by

$$R: \quad \mathbf{v}_1^{(2)} \leq \mathbf{v}_1^{(2)}(\mathbf{v}_1^{(1)}, \mathbf{v}_2^{(1)}, b) \leq \mathbf{v}_1^{(2)} + d\mathbf{v}_1^{(2)}.$$

For binary encounter we have in general

$$\rho = \frac{2m_2}{m_0} \mathbf{g} \cdot \boldsymbol{\kappa}, \quad (\text{A } 4)$$

where ρ ($= \mathbf{v}_1^{(2)} - \mathbf{v}_1^{(1)}$) is the velocity of the chosen molecule after collision relative to its velocity before collision, $m_0 = m_1 + m_2$ and $\boldsymbol{\kappa}$ is the unit vector in the direction of ρ . For rigid spheres we have in particular

$$b = \sigma_{12} \sin \psi, \quad gb db d\epsilon = \mathbf{g} \cdot \boldsymbol{\kappa} \sigma_{12}^2 d\Omega_{\kappa},$$

where $d\Omega_{\kappa} = \sin \psi d\psi d\epsilon$ is an elementary solid angle in the direction of $\boldsymbol{\kappa}$ and ψ is the angle between \mathbf{g} and $\boldsymbol{\kappa}$. Equation (A 3) then becomes

$$\chi_{12}^0(12) d\mathbf{v}_1^{(2)} = \int_R \int f_2^0(\mathbf{g} + \mathbf{v}_1^{(1)}) \sigma_{12}^2 \mathbf{g} \cdot \boldsymbol{\kappa} d\Omega_{\kappa} d\mathbf{g},$$

where the relative velocity \mathbf{g} has replaced $\mathbf{v}_2^{(1)}$ as integration variable. If $\boldsymbol{\kappa}$ is chosen as the new x -axis the region R can be described by two conditions: first, ρ must lie inside $d\Omega_{\kappa}$ and secondly $g_1 = \mathbf{g} \cdot \boldsymbol{\kappa}$ must lie within the limits

$$\frac{m_0}{2m_2} \rho \leq g_1 \leq \frac{m_0}{2m_2} (\rho + d\rho).$$

$$\text{Hence } \chi_{12}^0(12) d\mathbf{v}_1^{(2)} = d\Omega_\kappa \sigma_{12}^2 \int_{\frac{m_2}{2m_1}\rho}^{\frac{m_2}{2m_1}(\rho+d\rho)} g_1 dg_1 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_2^0(\mathbf{v}_1^{(1)} + \mathbf{g}) dg_2 dg_3,$$

where g_2 and g_3 are the two components of \mathbf{g} in the plane perpendicular to κ . After performing integration over g_2 and g_3 , one obtains

$$\chi_{12}^0(12) d\mathbf{v}_1^{(2)} = d\mathbf{v}_1^{(2)} n_2 \sigma_{12}^2 \left(\frac{m_2}{2\pi kT} \right)^{\frac{1}{2}} \left(\frac{m_0}{2m_2} \right)^2 \frac{\exp \left\{ -\frac{m_2}{2kT} \left[\frac{m_0}{2m_2} |\mathbf{v}_1^{(2)} - \mathbf{v}_1^{(1)}| + \mathbf{v}_1^{(1)} \cdot \kappa \right]^2 \right\}}{|\mathbf{v}_1^{(2)} - \mathbf{v}_1^{(1)}|}. \quad (\text{A } 5)$$

It may be verified that

$$\int \chi_{12}^0(12) d\mathbf{v}_1^{(2)} = \mu_{12}^0(1), \quad \int f^0(1) \chi_{12}^0(12) d\mathbf{v}_1^{(1)} = f_1^0(2) \mu_{12}^0(2),$$

as one would expect.

General force law

It has been mentioned in § 3 that for the general force law the classical cross-section diverges and the collision frequency μ also diverges due to the weak interaction of the distant molecules. Fortunately, we do not need to consider such weak interaction. Variation of r_0 will only result in the redistribution of the contributions from the successive terms in our final formula for the coefficient of self-diffusion (44).

Let $\phi(r)$ be the potential energy between two molecules expressed as a function of the distance r between their centres of gravity. We can take $\phi(r) = 0$ for $r > r_0$. The expression for $\mu_{12}(1)$ takes the same form as (A 1) with σ_{12} replaced by r_0 . It is the function χ which takes the actual force law inside the distance r_0 into account. We shall indicate here how $\chi_{12}^0(12)$ for the general force law can be calculated. As before, one has

$$\chi_{12}^0(12) d\mathbf{v}_1^{(2)} = \int_R \int f_2^0(\mathbf{v}_2^{(1)}) g b db d\epsilon d\mathbf{v}_2^{(1)}, \quad (\text{A } 6)$$

$$R: \quad \rho \leq \frac{2m_2}{m_0} \mathbf{g} \cdot \kappa \leq \rho + d\rho.$$

The differential equation of the orbit of relative motion in polar co-ordinates is

$$\frac{d\theta}{dr} = \frac{b}{r} \left[\left(1 - \frac{2\phi(r)}{Mg^2} \right) r^2 - b^2 \right]^{-\frac{1}{2}}, \quad (\text{A } 7)$$

where $M = \frac{m_1 m_2}{m_1 + m_2}$ is the reduced mass, r the radius vector and θ the corresponding polar angle. (The polar axis extends from the centre of the chosen molecule along the direction of $-\mathbf{g}$.)

At the moment of nearest approach, the polar co-ordinates of the incoming molecule being R and ψ , one has $dr/d\theta = 0$. Hence

$$\frac{2\phi(R)}{Mg^2} = 1 - \frac{b^2}{R^2}, \quad (\text{A } 8)$$

from which R as a function of b and g can in principle be found. At the moment when the incoming molecule first enters the sphere of the influence, its polar co-ordinates are r_0 and θ_0 with $\theta_0 = \sin^{-1}(b/r_0)$. From (A 6), regarding θ as a function of r and integrating between R and r_0 , one has

$$\psi = \theta_0 + \int_{R(b, \theta)}^{r_0} \frac{b}{r} \left[\left(1 - \frac{2\phi(r)}{Mg^2} \right) r^2 - b^2 \right]^{-\frac{1}{2}} dr = \psi(b, g), \quad (\text{A } 9)$$

from which b as a function of g and ψ can in principle be found. Equation (A 6) becomes then

$$\begin{aligned} \chi_{12}^0(12) d\mathbf{v}_1^{(2)} &= \iint_R f_2(\mathbf{v}_2^{(1)}) g \frac{b(g, \psi)}{\sin \psi} \left| \frac{\partial b}{\partial \psi} \right| \sin \psi d\psi d\epsilon d\mathbf{g} \\ &= \rho^2 d\Omega_\kappa \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{\frac{m_0}{2m_a} \rho}^{\frac{m_0}{2m_a} (\rho + d\rho)} f_2(\mathbf{g} + \mathbf{v}_1^{(1)}) \frac{g}{\rho^2} \frac{b(g, \psi)}{\sin \psi} \left| \frac{\partial b}{\partial \psi} \right| dg_1 dg_2 dg_3, \end{aligned} \quad (\text{A } 10)$$

where ψ inside the integral is expressed by $\psi = \cos^{-1} \frac{g_1}{g}$.

APPENDIX B. CALCULATION OF α_1 AND β'_1 FOR NON-UNIFORM GAS MIXTURES

The assumption that the physical parameters λ_i vary only smoothly will be made in the calculation. One can always express λ_i in the vicinity of a given point \mathbf{x} by means of Taylor's expansion. For the present purpose we shall neglect the presence of a velocity gradient and assume a constant mass-motion velocity in the region in which $\psi(\Delta\mathbf{x}, \Delta t)$ is appreciable. The velocity distribution functions of the two kinds of molecules at \mathbf{x}, t normalized to their respective local number densities $n_1(\mathbf{x}, t)$ and $n_2(\mathbf{x}, t)$ are denoted by $f_1(\xi_1, \mathbf{x}, t)$ and $f_2(\xi_2, \mathbf{x}, t)$. The following abbreviations will be used hereafter:

$$\left. \begin{aligned} f_1(2)_3 &= f_1(\xi_1^{(2)}, \mathbf{x}^{(3)}, t_3), \quad \mu_1(2)_3 = \mu(\xi_1^{(2)}, \mathbf{x}^{(3)}, t_3), \quad n_1(2) = n_1(\mathbf{x}^{(3)}, t_2), \\ \chi_1(23)_4 &= \chi(\xi_1^{(2)}, \xi_1^{(3)}, \chi^{(4)}, t_4), \quad D_2^{(3)} = \xi_2^{(3)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}} + \frac{\partial}{\partial t_1}, \\ \frac{\partial}{\partial \mathbf{x}^{(1)}} &= \sum_i \frac{\partial \lambda_i}{\partial \mathbf{x}^{(1)}} \frac{\partial}{\partial \lambda_i}, \quad \frac{\partial}{\partial t_1} = \sum_i \frac{\partial \lambda_i}{\partial t_1} \frac{\partial}{\partial \lambda_i} \quad (\lambda_i = n_1, n_2, T). \end{aligned} \right\} \quad (\text{B } 1)$$

Further, if the pressure is taken to be constant during diffusion and it is supposed that no external force field is present, one finds

$$\left(\mathbf{u}_2 \frac{\partial}{\partial \mathbf{x}^{(1)}} + \frac{\partial}{\partial t_1} \right) \lambda_i = 0 \quad \text{or} \quad D_2^{(3)} = \mathbf{v}_2^{(3)} \cdot \frac{\partial}{\partial \mathbf{x}^{(1)}}, \quad (\text{B } 2)$$

where $\mathbf{v}_2^{(3)} (= \xi_2^{(3)} - \mathbf{u})$ is the peculiar velocity of a molecule of the second kind on its i th flight relative to the local mass motion velocity \mathbf{u} .

Both μ and χ at any later stage of the motion of the molecule can be expressed in terms of their values at the starting-point:

$$\left. \begin{aligned} \mu_1(i)_j &= \mu_1(i)_1 + \sum_{l=1}^{j-1} \tau_{l, l+1} D_1^{(l)} \mu_1(i)_1 + \dots, \\ \chi_1(i, i+1)_j &= \chi_1(i, i+1)_1 + \sum_{l=1}^{j-1} \tau_{l, l+1} D_1^{(l)} \chi_1(i, i+1)_1 + \dots \end{aligned} \right\} \quad (\text{B } 3)$$

By means of (B 3), $P_N(1, 2, \dots, N)$ in (39) can be expressed in terms of the values of quantities at the starting-point $\mathbf{x}^{(1)}, t_1$, neglecting gradients of the second and higher order. The mean time of flight, mean displacement and mean square displacement relative to a frame moving with \mathbf{u} can be found by simple calculations. The mean time of flight is found to depend on the direction of the velocity in this flight, but as the difference can at most be linear in the gradients, it will not affect our final result for α_i, β'_i which is obtained by using the average time of flight for all directions in each flight. Further, though such times averaged over all directions are different for individual flights, the difference can again be ignored for the same reason.

As it is understood in the following that the chosen molecule is of the first kind and quantities pertaining to any later stage of its journey are always expressed in terms of those at the starting-point $\mathbf{x}^{(1)}, t_1$, we shall drop subscripts attached to f, μ and χ and write $f(1)$ for $f_1(1)_1$, $\mu(i)$ for $\mu_1(i)_1$ and $\chi(i, i+1)$ for $\chi_1(i, i+1)_1$. With this simplification one writes, for the mean time of the i th flight over all directions,

$$\langle \tau_{i, i+1} \rangle = \frac{1}{\mu^0(1)} + O\left(\frac{\partial \lambda_i}{\partial \mathbf{x}}\right), \quad (\text{B } 4)$$

where
$$\overline{\mu^0(1)} = \frac{1}{n_1} \int f^0(1) \mu^0(1) d\mathbf{v}_1^{(1)},$$

and the total time of flight for N successive flights is

$$\Delta t = \left\langle \sum_{i=1}^N \tau_{i, i+1} \right\rangle = \frac{N}{\mu^0(1)} + O\left(\frac{\partial \lambda_i}{\partial \mathbf{x}}\right). \quad (\text{B } 5)$$

The successive mean displacements relative to \mathbf{u} are

$$\left. \begin{aligned} \langle \mathbf{l}^{(1)} \rangle &= \langle \mathbf{v}_1^{(1)} \tau_{12} \rangle = \frac{\iint P_1(1) \tau_{12} \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} d\tau_{12}}{\iint P_1(1) d\mathbf{v}_1^{(1)} d\tau_{12}}, \\ \langle \mathbf{l}^{(2)} \rangle &= \langle \mathbf{v}_1^{(2)} \tau_{23} \rangle = \frac{\iint P_2(12) \mathbf{v}_1^{(2)} \tau_{23} d\mathbf{v}_1^{(1)} d\mathbf{v}_1^{(2)} d\tau_{12} d\tau_{23}}{\iint P_2(12) d\mathbf{v}_1^{(1)} d\mathbf{v}_1^{(2)} d\tau_{12} d\tau_{23}}, \dots, \end{aligned} \right\} \quad (\text{B } 6)$$

where

$$\begin{aligned} P_1(1) &= (\mu(1) + D_1^{(1)} f(1)) \left\{ \exp \left[- \int_0^{\tau_{12}} (\mu(1) + \tau_{12} D_1^{(1)} \mu^0(1)) d\tau \right] \right\} (\mu(1) + \tau_{12} D_1^{(1)} \mu^0(1)) \\ &= f^0(1) (\mu^0(1))^2 \exp [-\mu^0(1) \tau_{12}] \left\{ 1 + \frac{D_1^{(1)} f^0(1)}{\mu^0(1) f^0(1)} - \frac{\tau_{12}^2}{2} D_1^{(1)} \mu^0(1) + \frac{\tau_{12} D_1^{(1)} \mu^0(1)}{\mu^0(1)} \right. \\ &\quad \left. + \frac{f'(1)}{f^0(1)} + \frac{2\mu'(1)}{\mu^0(1)} - \mu'(1) \tau_{12} \right\}, \quad (\text{B } 7) \end{aligned}$$

.....

neglecting terms involving the second or higher powers of $\partial \lambda_i / \partial \mathbf{x}$.

With the help of (B 7), equation (B 6) becomes

$$\begin{aligned}\langle \mathbf{I}^{(1)} \rangle &= \frac{1}{n\mu^0(1)} \left\{ \int f'(1) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} + \int D_1^{(1)} \left(\frac{f^0(1)}{\mu^0(1)} \right) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} \right\}, \\ \langle \mathbf{I}^{(2)} \rangle &= \frac{1}{n\mu^0(1)} \left\{ \int f'(1) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} + \int D_1^{(1)} \left(\frac{f^0(1)}{\mu^0(1)} \right) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} \right. \\ &\quad \left. + \int \int D_1^{(1)} \left(f^0(1) \frac{\chi^0(12)}{\mu^0(1)\mu^0(2)} \right) \mathbf{v}_1^{(2)} d\mathbf{v}_1^{(1)} d\mathbf{v}_1^{(2)} \right\}, \\ &\dots\dots\end{aligned}$$

Hence the total mean displacement $\overline{(\Delta \mathbf{x})}$ is

$$\begin{aligned}\overline{(\Delta \mathbf{x})} &= \sum_{i=1}^N \langle \mathbf{I}^{(i)} \rangle = \frac{N}{n_1\mu^0(1)} \left\{ \int f'(1) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} + \int D_1^{(1)} \left(\frac{f^0(1)}{\mu^0(1)} \right) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} \right. \\ &\quad \left. + \left(\frac{N-1}{N} \right) \int \int D_1^{(1)} \left(f^0(1) \frac{\chi^0(12)}{\mu^0(1)\mu^0(2)} \right) \mathbf{v}_1^{(2)} d\mathbf{v}_1^{(1)} d\mathbf{v}_1^{(2)} + \dots \right\}. \quad (\text{B } 8)\end{aligned}$$

The mean square displacement when quantities involving the gradients are neglected is found to be the same as in the equilibrium state. The required expressions for α_1 and β'_1 follow from (B 5), (40) and (B 8) for large N :

$$\left. \begin{aligned}\alpha_1 &= \left\{ \int \frac{f^0(1)}{n_1} \frac{(\mathbf{v}_1^{(1)})^2}{\mu^0(1)} d\mathbf{v}_1^{(1)} + \int \int \frac{f^0(1)}{n_1} \frac{\chi^0(12)}{\mu^0(1)\mu^0(2)} \mathbf{v}_1^{(1)} \cdot \mathbf{v}_1^{(2)} d\mathbf{v}_1^{(1)} d\mathbf{v}_1^{(2)} + \dots \right\}, \\ \beta'_1 &= \frac{1}{n_1} \left\{ \int f'(1) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} + \int D_1^{(1)} \left(\frac{f^0(1)}{\mu^0(1)} \right) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} \right. \\ &\quad \left. + \int \int D_1^{(1)} \left(f^0(1) \frac{\chi^0(12)}{\mu^0(1)\mu^0(2)} \right) \mathbf{v}_1^{(2)} d\mathbf{v}_1^{(1)} d\mathbf{v}_1^{(2)} + \dots \right\}.\end{aligned} \right\} \quad (\text{B } 9)$$

With the help of (B 2), β'_1 can be reduced to the form

$$\beta'_1 = \frac{1}{n_1} \int f'(1) \mathbf{v}_1^{(1)} d\mathbf{v}_1^{(1)} + \frac{1}{3n_1} \frac{\partial}{\partial \mathbf{x}^{(1)}} (n_1 \alpha_1) = \mathbf{u}'_1 + \frac{1}{3n_1} \frac{\partial}{\partial \mathbf{x}^{(1)}} (n_1 \alpha_1). \quad (\text{B } 10)$$

Substituting (B 10) in (13), an identity is obtained which proves the statement made at the beginning of § 5.

APPENDIX C. EVALUATION OF THE COEFFICIENT OF SELF-DIFFUSION IN GASES

To evaluate the coefficient of self-diffusion in a simple gas as given in (44), we need μ and χ which are given in (A 1) and (A 5) of appendix A for the rigid spherical model. For a simple gas $\mu_{12}^0(1)$ and $\chi_{12}^0(12)$ reduce to the form

$$\left. \begin{aligned}\mu(1) &= n\pi\sigma^2 \left(\frac{2kT}{\pi m} \right)^{\frac{1}{2}} E(y_1), \\ \chi(12) &= n\sigma^2 \left(\frac{m}{2\pi kT} \right)^{\frac{1}{2}} \frac{1}{\rho_{12}} \exp \left\{ - \left[y_2^2 - \left(\frac{y_1 y_2}{\rho_{12}} \sin \theta_{12} \right)^2 \right] \right\},\end{aligned} \right\} \quad (\text{C } 1)$$

$$\left. \begin{aligned}\text{where } y_1 &= \sqrt{\left(\frac{m}{2kT} \right) \mathbf{v}^{(1)}}, \quad y_2 = \sqrt{\left(\frac{m}{2kT} \right) \mathbf{v}^{(2)}}, \\ \rho_{12}^2 &= y_1^2 + y_2^2 - 2y_1 y_2 \cos \theta_{12}, \quad E(y_1) = e^{-y_1^2} + \left(2y_1 + \frac{1}{y_1} \right) \int_0^{y_1} e^{-x^2} dx.\end{aligned} \right\} \quad (\text{C } 2)$$

θ_{12} is the angle between \mathbf{y}_1 and \mathbf{y}_2 , n , m and σ are the number density, the molecular mass and diameter respectively.

Substituting (C1) in (44) one obtains

$$D_s = \frac{4}{3\pi\sigma^2 n} \left(\frac{2kT}{m} \right)^{\frac{1}{2}} \{c_0 + c_1 + c_2 + \dots\}, \quad (\text{C3})$$

where

$$\left. \begin{aligned} c_0 &= \int_0^\infty \frac{e^{-v_1^2} dy_1}{E(y_1)}, \\ c_1 &= 2 \int_0^\infty \int_0^\infty \int_0^\pi \frac{e^{-(v_1^2+v_2^2)}}{E(y_1)E(y_2)} \frac{e^{(v_1 v_2 \sin \theta_{12}/\rho_{12})^2}}{\rho_{12}} (y_1 y_2)^3 \sin \theta_{12} \cos \theta_{12} d\theta_{12} dy_1 dy_2, \\ c_2 &= 4 \int_0^\infty \int_0^\infty \int_0^\pi \int_0^\pi \frac{e^{-(v_1^2+v_2^2+v_3^2)}}{E(y_1)E(y_2)E(y_3)} \left\{ \exp \left[\left(\frac{y_1 y_2 \sin \theta_{21}}{\rho_{12}} \right)^2 + \left(\frac{y_2 y_3 \sin \theta_{23}}{\rho_{23}} \right)^2 \right] \right. \\ &\quad \times \left. \left(\frac{y_1^3 y_2^3 y_3^3}{\rho_{12} \rho_{23}} \right) \sin \theta_{21} \sin \theta_{23} \cos \theta_{13} d\theta_{21} d\theta_{23} dy_1 dy_2 dy_3, \right. \\ &\quad \dots \end{aligned} \right\} \quad (\text{C4})$$

The term c_0 has been evaluated accurately by means of the table of functions provided by P. Herberg to be 0.2093. To estimate the values of the rest we make the following approximation: the velocity spread of the Maxwellian velocity distribution is replaced inside the integral by a distribution of constant average speed with random orientation, i.e. $\left(\frac{m}{2\pi kT} \right)^{\frac{3}{2}} \exp \left[-\frac{m}{2kT} v^{(1)2} \right] dv^{(1)}$ is replaced by $\frac{1}{2\pi \bar{v}^{(1)}} \delta(|v^{(1)}|^2 - \bar{v}^{(1)2}) dv^{(1)}$, where $\bar{v}^{(1)} = 2 \left(\frac{2kT}{\pi m} \right)^{\frac{1}{2}}$, or $e^{-v_1^2}$ is replaced by $\frac{\pi}{4} \delta(|y_1|^2 - \bar{y}_1^2)$ with $\bar{y}_1 = \frac{2}{\sqrt{\pi}}$.

By introducing such δ -functions, c_1 and c_2 can be worked out after short calculations to be

$$c_1 = \frac{\pi^{\frac{3}{2}}}{8E^2(2/\sqrt{\pi})} \left\{ 1 + \left(1 - \frac{\pi}{4} \right) e^{4/\pi} \Phi \left(\frac{2}{\sqrt{\pi}} \right) \right\} = 0.15,$$

$$c_2 = \frac{\pi^{\frac{3}{2}}}{256E^3(2/\sqrt{\pi})} \left\{ \left[1 + \left(1 - \frac{\pi}{4} \right) e^{4/\pi} \Phi \left(\frac{2}{\sqrt{\pi}} \right) \right]^2 + \left[e^{4/\pi} - \frac{\sqrt{\pi}}{2} \int_0^{2/\sqrt{\pi}} e^{t^2} dt \right]^2 \right\} = 0.07,$$

where $\Phi \left(\frac{2}{\sqrt{\pi}} \right) = \frac{2}{\sqrt{\pi}} \int_0^{2/\sqrt{\pi}} e^{-t^2} dt = 0.89$, $E \left(\frac{2}{\sqrt{\pi}} \right) = 2.76$.

Up to the third term the value of D_s is

$$D_s = \frac{4}{3\pi n \sigma^2} \left(\frac{2kT}{m} \right)^{\frac{1}{2}} (0.43) = 0.81 \frac{1}{n\pi\sigma^2} \left(\frac{kT}{m} \right)^{\frac{1}{2}}. \quad (\text{C5})$$

This estimation serves, on the one hand, to show the rapid convergence of the series of constants c_0, c_1, \dots and on the other to compare with the expression given by Chapman and Pidduck,

$$D_s = (1.019) \frac{3}{8n\sigma^2} \left(\frac{kT}{\pi m} \right)^{\frac{1}{2}} = 0.779 \frac{1}{n\pi\sigma^2} \left(\frac{kT}{m} \right)^{\frac{1}{2}}. \quad (\text{C6})$$

REFERENCES

- Chandrasekhar, S. 1943 *Rev. Mod. Phys.* **15**, 1.
 Chapman, S. & Cowling, T. G. 1939 *The mathematical theory of non-uniform gases*. Cambridge University Press.
 Einstein, A. 1905 *Ann. Phys., Lpz.*, **17**, 549.
 Einstein, A. 1906 *Ann. Phys., Lpz.*, **19**, 371.
 Fürth, R. 1942 *Proc. Roy. Soc. A*, **179**, 461.
 Jeans, J. H. 1925 *Dynamical theory of gases*. Cambridge University Press.
 Kennard, E. H. 1938 *Kinetic theory of gases*. New York: McGraw-Hill Book Co.
 Rosenberg, P. 1942 *Phys. Rev.* **61**, 528.
 Wang, M. C. & Uhlenbeck, G. E. 1945 *Rev. Mod. Phys.* **15**, 1.

Eddy diffusion of water vapour and heat near the ground

BY F. PASQUILL, *Meteorological Office, London and School of Agriculture,
University of Cambridge*

(Communicated by Sir Geoffrey Taylor, F.R.S.—Received 17 December 1948)

An experimental study has been made of the factors involved in the turbulent transport of water vapour and heat in the lowest layer of the atmosphere over well-exposed level grassland. Measurements were made over periods of 1 hr. of the water loss from isolated but otherwise naturally exposed sections of the surface layers of the soil and quantitative arguments advanced for adopting them as a reasonable approximation to the true evaporation loss from the ground surface. The incoming and reflected components of solar radiation, the temperature distribution in the soil down to 16 in. and the vertical profiles of temperature, humidity and wind speed in the air up to a height of 2 m. were observed at the same time, and samples taken to provide necessary data on the physical properties of the soil. The net flux of long-wave radiation was computed from the temperature and humidity structure of the atmosphere as given by the present low-level measurements and routine upper-air soundings. The data prescribe the vertical turbulent flux and the vertical gradients of the water vapour and heat content of the air, from which may be evaluated the vertical components of the eddy diffusivities for water vapour and heat (K_v and K_H) as customarily defined.

In the absence of thermal stratification of the surface air layers K_v is shown to be identical with the eddy diffusivity for momentum (K_m) defined by the explicitly established logarithmic law relating the aerodynamic drag and vertical wind shear over a rough surface. The modification of K_v by unstable and stable thermal stratifications and the rapid decrease of stability influence as the ground surface is approached are both quantitatively demonstrated, and a unique relation between parameters involving K_v , the vertical wind shear and the vertical temperature gradient is indicated. No completely satisfactory *a priori* explanation can as yet be given for the latter relation, though in unstable conditions K_v is found to be identical with K_m computed from a recent wind-profile law which does not involve the temperature gradient explicitly and has only been established in functional form. Direct comparison of K_v and K_H reveals a reasonable approach to equality in stable conditions but shows that the latter coefficient is systematically and substantially the greater in unstable conditions. The latter feature is qualitatively in keeping with recent trends in the theoretical concepts of turbulent transport.

The bearing of the results on the problem of indirectly evaluating natural land evaporation is briefly discussed and attention drawn to the implied superiority of the present 'hydrodynamical' approach over the classical 'heat-balance' method.

1. INTRODUCTION

The earliest considerations of the effect of the turbulent structure of the lower atmosphere, initiated by Taylor (1915, 1917), clearly demonstrated the existence of a diffusive action much more powerful than any which could be attributed to molecular

motion. From data then available on the vertical distribution of wind velocity and temperature in the atmosphere Taylor deduced 'eddy diffusivities' (analogous to the kinematic viscosity and thermal diffusivity in molecular transport) of the order of 10^3 to 10^5 cm.²/sec. and confirmed in a broad sense his theoretical deduction that the eddy transport of conserved properties of the atmosphere may be prescribed by an eddy diffusivity which is independent of the property undergoing transport. Hydrographic observations by Jacobsen (see Taylor 1931) indicated, however, that the latter feature did not necessarily hold in a fluid possessing a marked density stratification and in which gravitational forces were thus of dominant importance.

The precise magnitudes of the eddy diffusivities appropriate to the transport of momentum, matter and heat, and their dependence on other physical factors, still occupy a central place of interest in problems of turbulent diffusion. Much accurate information has arisen from laboratory studies of aerodynamic drag, evaporation and heat transfer in turbulent flow through pipes and over flat plates, and it is noteworthy that in the case of aerodynamically smooth surfaces there is good evidence (Pasquill 1943) for close numerical equality in the eddy diffusivities for momentum, vapour and heat, except for small differences which may reasonably be ascribed to different magnitudes of the appropriate molecular diffusivities. Data for atmospheric flow are necessarily less precise and are complicated by the natural roughness of the earth's surface and the frequently marked thermal stratification of the surface layers of the air. In the absence of the latter feature treatments by Sutton (1934, 1947) and Calder (1949), assuming identity in the transport of momentum and matter, have made substantial advances in our understanding of the spread of smoke and vapour introduced artificially into the atmosphere. On the other hand, the present knowledge of the processes of diffusion of the water vapour and heat content of the atmosphere is still based in the main upon more or less reasonable conjecture untested as yet by critical observational data. Furthermore, a recent notable contribution to the theoretical side of the problem (Priestley & Swinbank 1947) introduces in the heat-flux equation a 'buoyancy' term which had not been explicitly recognized previously, though the importance of gravitational forces in turbulent processes had already been noted by Taylor (1931 and unpublished), and in broader implication implies the possibility of there being different values of the eddy diffusivity for the turbulent transport of different properties in a thermally stratified atmosphere.

An observational study with quantitative bearing on some of the above features is presented and discussed in this paper. The study is restricted to a shallow surface layer of the atmosphere and to flow over a level well-exposed surface of small aerodynamic roughness, but it covers a substantial range in the magnitude of the wind velocity and the degree of thermal stratification of the atmosphere and provides some appreciation of the problem of indirectly evaluating natural evaporation.

2. OBSERVATIONAL ASPECTS

The observational problem consisted of obtaining values of the vertical flux of water vapour, in effect the natural rate of evaporation from the ground, the vertical flux of heat by turbulent transport and the vertical gradients of humidity and temperature. From such data the magnitudes of the eddy diffusivities for water vapour

and heat are directly obtainable, while from measurements of the vertical profile of wind speed and appeal to aerodynamic laws the eddy diffusivity for momentum may also be evaluated in certain circumstances. The measurements were made on a clayland pasture on the University Farm at Cambridge during spells of fine weather in March 1948. For the wind directions then obtaining this site is suitably level and unobstructed for the establishment of horizontal homogeneity in the structure of the air up to a height of 2 m. Upwind of the instruments there was always at least 150 yd. of level uniform terrain (see Deacon 1949, on the distance required for the setting up of a homogeneous wind profile). The nearest substantial variation in level occurs beyond 150 yd. to the south-east where the ground slopes gently into a small valley, while the nearest obstacle of important size is a close about 50 ft. high covering a front of 250 yd. at a distance of 500 yd. to the south-south-west. Bare soil was visible over a considerable proportion of the pasture and the patchy grass cover was mainly 2 cm. long with occasional withered tufts up to 5 to 7 cm. and showed no appreciable change in length over the whole period of observation.

The essential features of the assembly of instruments are illustrated in figure 1. The vertical profiles of air temperature and humidity up to a height of 2 m. were explored over individual periods of 1 hr. with a portable distant-indicating thermocouple psychrometer apparatus developed specially for the purpose and described in detail elsewhere (Pasquill 1949), while the corresponding profiles of wind speed were observed with sensitive cup anemometers which are a development of a type described by Sheppard (1940). Masts supporting the instruments at heights of 200, 150, 100, 50, 37.5 and 25 cm. were set up in line across wind. During each 3 min. period the differences in dry-bulb and wet-bulb temperature between the 200 cm. level and each other level in turn were read to 0.01°F on a suitable indicating galvanometer, and the absolute values of the temperatures at the 200 cm. level were read separately to 0.1°F . In the earlier observations this procedure was maintained throughout the whole period of 1 hr., the anemometers being continuously in operation over the same period. At the end of the run the psychrometer mast was swung into a horizontal position across wind and five sets of readings taken so as to provide zero corrections. At the same time all anemometers were mounted at a height of 2 m. and compared over a period of 20 min. or so. In these control observations there was sometimes reason to suspect that the apparent inequalities in psychrometer performance were partly spurious and due to lateral wind swings around the closely disposed instruments, causing a given psychrometer to receive air which had been in contact with the irradiated surface of an adjacent instrument. A modified control procedure was accordingly adopted in all subsequent observations. In this the observation period was divided into two equal parts with a 12 min. interval, and for the second part the psychrometers and anemometers were systematically interchanged, the 200 cm. instrument with that at 100 cm., 150 cm. with 37.5 cm. and 50 cm. with 25 cm. On the basis of laboratory measurements of thermal lag it was known that the period of 12 min. was adequate for the adjustment of the interchanged psychrometers to their new environments. This procedure complicated the observational and analytical work but on the whole was thought to provide a more dependable instrumental control.

For the evaluation of the vertical flux of heat it was necessary to measure or estimate the magnitudes of the remaining components in the balance of heat-exchange processes. The direct and diffuse solar radiation on a horizontal surface and the component reflected from the ground were measured by a solarimeter calibrated at Kew Observatory, one reading of each component being taken in each 3 min. period. These measurements were made only with practically clear sky, so that individual readings did not show the pronounced variation associated with scattered or broken cloud, and the frequency of reading adopted could thus be expected to yield representative mean values.

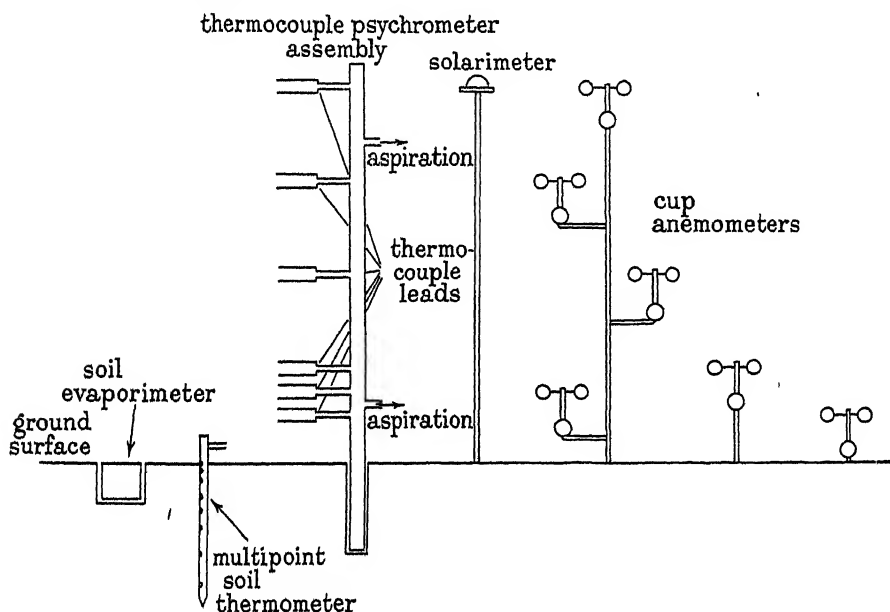


FIGURE 1. Schematic diagram of array of instruments.

Soil temperatures at various depths were measured by a distant-indicating multi-point thermometer consisting of a non-conducting cylindrical stem with copper strips set in various positions flush with the outer surface. Fine thermocouple wires, enclosed in the stem, terminate in single junctions soldered into each copper strip. The essential principle of the instrument is that it may be installed in the soil, merely by driving it in, with a minimum disturbance of soil conditions together with the maintenance of adequate contact between the soil and the temperature-sensitive surfaces. The temperature differences between the thermocouple at a depth of 16 in. and those at 12, 8, 6, 4, 2, 1, $\frac{1}{2}$ and approximately 0 in. were read to 0.05°F during each 3 min. period, and a reading corresponding to the difference between the 16 in. element and an element immersed in a water-bath at known temperature was also taken so as to specify the absolute values of the temperatures.

In order subsequently to obtain data on the thermal properties of the soil samples were taken with an 'undisturbed core' sampler consisting of a steel tube with a cutting rim tapered on the outside and inset slightly on the inside, the cutting diameter being 4.5 cm. On driving the sampler vertically into the ground compression

and distortion of the soil occurs outside the tube, while a relatively undisturbed core of soil is left inside the tube, from which it is quickly removed, cut into suitable sections and weighed immediately. These samples were taken down to a depth of 8 in., at a single point about a yard away from the soil thermometer, on three occasions corresponding to the general periods when heat-balance observations were made.

Evaluation of the long-wave radiation term requires, among other items, a knowledge of the radiative temperature of the ground. It has been shown by Robinson (1947) that to a good degree of approximation the radiative temperature of ground covered by short grass is indicated by a spirit thermometer lying on the ground (actually a Meteorological Office pattern grass minimum thermometer), even when the sun is at appreciable elevation. Single readings were taken in each of the 3 min. periods, this again being considered of adequate frequency in clear weather when violent oscillations of the indicated temperature did not occur.

Measurements of the natural rate of evaporation from the ground were attempted by means of soil evaporimeters of simple design. Undisturbed cores of soil 4 in. deep were extracted by means of a sampler similar to that already described but of cutting diameter $3\frac{1}{8}$ in. These were inserted in close-fitting pots made of Tufnol tube of internal diameter 4 in. and wall thickness $\frac{1}{4}$ in., so that the soil surface was flush with the rim. Corresponding but slightly larger containers were embedded in the ground with the rims flush with the surrounding soil surface, so forming clean watertight receptacles for the soil containers. The effect was thus to isolate sections of the surface layer of the ground from surrounding and underlying soil without markedly disturbing the conditions existing therein, while leaving the actual surfaces exposed as in the natural state. These soil evaporimeters were set out about 15 to 20 yd. upwind of the temperature and wind instruments and weighed at the beginning and end of the observation period of 1 hr. on an indicating balance with an accuracy better than 0.03 g. The losses ranged from 0.1 to 2.5 g., and corrections (usually about 5 %) were applied for the abnormal exposure during the weighing process.

3. THE INTERPRETATION OF THE EVAPORATION MEASUREMENTS

The direct determination of the true rate of evaporation by the preceding method is complicated in particular by the unavoidable isolation of the test soil from the underlying layers of soil, though it was thought that it might be possible to choose circumstances in which the effect of this limitation could be neglected. The site of the present measurements is on a deep bed of clay, the top layer of which normally dries out fairly progressively from the surface downward during the spring and summer. It follows that there may be certain stages in the drying history of such ground when isolation of the soil at *some shallow depth* would not appreciably affect the rate of loss of water from the surface *for some period*. Such a stage possibly occurs in the early spring before drying out has proceeded to any great depth and before root growth has penetrated deeply, thus possibly drawing water from lower strata.

The optimum dimensions of isolated soil cores consistent with maximum depth of isolation, capacity and accuracy of balances available and anticipated order of

the rate of evaporation, were in the region of 4 in. diameter and depth, and these dimensions were accordingly adopted for the main soil evaporimeters. Two features suggested qualitatively that the isolation imposed at a depth of 4 in. was probably unimportant in the present measurements. In the course of extracting soil cores it was observed that the root growth rarely penetrated below 3 in., so that transpiration was probably unaffected. Furthermore, it was unmistakably evident that appreciable drying had occurred in the top 2 in. of previously undisturbed ground, whereas there was no visual indication of drying having occurred at lower levels. However, a quantitative test was clearly necessary, and an approach towards this was provided in the following fashion. Additional soil pots were made to take soil cores of the same surface area but of depths 2 and 3 in. respectively, and on all occasions one or both of these was used in conjunction with the normal 4 in. evaporimeters. If the vertical isolation exerted substantial influence on the water loss from the core it would be expected that the effects would show up to an increasing degree with the decrease of isolation depth and with the passage of time.

TABLE 1. EVAPORATION DATA

group and date	period of evapora- tion (min.)	obs. no.	G.M.T.*	relative loss from evaporimeters							actual loss in g. from no. 5	
				2 in.	3 in.	no. 1	no. 2	no. 3	no. 4	no. 5		
10. iii. 48	A	75	6	10.33	1.62	1.21	1.38	2.19	1.40	1.33	1.00	0.82
		75	7	12.30	1.69	1.14	1.17	1.62	1.24	1.25	1.00	1.00
		75	8	15.30	1.69	1.03	1.33	1.49	1.27	1.39	1.00	0.75
		75	9	17.30	2.92	2.17	1.83	2.83	2.17	4.17	1.00	0.12
		75	10	19.30	1.82	1.00	1.29	1.41	1.29	1.82	1.00	0.17
18. iii. 48	B	75	11	12.35	0.75	0.75	0.79	0.85	0.83	0.89	1.00	2.48
		75	12	14.30	0.74	0.74	0.77	0.88	0.88	0.82	1.00	1.60
		75	13	16.34	0.73	0.68	0.85	0.80	0.89	0.68	1.00	0.99
		75	14	19.32	0.53	0.61	0.66	0.84	0.76	0.68	1.00	0.38
24. iii. 48	C	70	15	09.38	0.78	0.92	1.00	0.95	—	—	—	1.18
		70	16	11.30	0.69	0.87	1.00	0.89	—	—	—	1.47
		70	17	14.30	0.68	0.84	1.00	0.82	—	—	—	1.45
		70	18	16.30	0.77	0.87	1.00	1.06	—	—	—	0.88
		70	19	18.30	0.63	0.77	1.00	0.67	—	—	—	0.30
25. iii. 48	C'	70	20	10.30	0.71	0.91	1.00	0.84	—	—	—	1.14
		70	21	12.35	0.76	0.92	1.00	0.99	—	—	—	1.43
		70	22	15.30	0.80	0.84	1.00	0.87	—	—	—	1.28
		70	23	18.30	0.62	0.85	1.00	1.00	—	—	—	0.26
28. iii. 48	D	70	24	10.30	0.81	0.89	1.00	0.91	—	—	—	1.52
		70	25	12.33	0.76	0.83	1.00	0.86	—	—	—	1.62
		70	26	14.42	0.82	0.97	1.00	0.97	—	—	—	1.10
		70	27	16.30	0.64	0.91	1.00	0.94	—	—	—	0.64
		70	28	19.00	0.11	0.67	1.00	1.00	—	—	—	0.09
29. iii. 48	D'	70	29	11.39	0.94	1.08	1.00	1.00	—	—	—	1.35
		70	30	14.03	0.79	0.98	1.00	1.00	—	—	—	0.85

NOTE. Evaporimeters nos. 1 to 6 are of depth 4 in.

* Refers to midtime of observation period of approximately 60 min.

Details of the water losses observed during each observation period are listed in table 1. The observations have been grouped into sets in which the same soil cores were employed for a number of consecutive observations, these sets being designated by letters *A*, *B*, *C* and *D*, a subdivision, indicated by a dashed letter, being made when a set of soil cores were employed for 2 days in succession (*C'* and *D'*). The figures include the actual loss from one of the evaporimeters (no. 5) and the relative losses from the remainder. For group *A* evaporimeters nos. 1 and 2 and the 2 and 3 in. auxiliary evaporimeters were installed the day before, whereas nos. 3, 4 and 5 had been in position for 8 days, during which period each had lost about 25 g., which corresponded to approximately 10 % of the original water content. For group *C* the installation was carried out 2 days preceding, while for groups *B* and *D* it was carried out on the same morning.

Group *A*, in which the maximum number of evaporimeters was employed, provided the most critical test of performance in view of the wide range of evaporimeter age also involved. In any one observation the losses indicated by individual evaporimeters vary substantially, though in the whole group of observations there is apparently a fairly consistent relation between individual evaporimeters. We may therefore conveniently examine these results in terms of the mean relative losses for the whole group, and these are reproduced below:

evaporimeter			no. 1	no. 2	no. 3	no. 4	no. 5
evaporimeter depth	2 in.	3 in.	4 in.	4 in.	4 in.	4 in.	4 in.
evaporimeter 'age'	1 day				7 days		
mean relative loss	1.9	1.3	1.4	1.9	1.5	2.0	1.0

From these figures no systematic effect of either evaporimeter depth or evaporimeter age is evident, and it seems legitimate to conclude that the differences exhibited are primarily due to a genuine point-to-point variation in evaporation, arising from heterogeneity in soil moisture, vegetation cover and local exposure. From inspection of the rest of the observations, for which in each group evaporimeter age was constant, there is again no pronounced indication of a systematic effect of evaporimeter depth. On the argument put forward these features strongly suggest that in the soil situation then existing the soil isolation imposed had no important effect on the rate of loss of water from the surface, and that the losses observed with the 4 in. evaporimeters may be assumed to be a reasonable approximation to the true rate of evaporation.*

4. REDUCTION OF DATA ON HUMIDITY, TEMPERATURE AND WIND PROFILES

Mean values of wind speed, temperature and computed absolute humidity at each of the six heights are given for each observation in table 2.† Each figure represents a mean value over a period of 1 hr., and in the case of temperature and humidity

* The writer was aware, prior to designing these measurements, that a method of measuring natural rate of evaporation, using plastic evaporimeter tubes extending to various depths, was in course of development by W. C. Swinbank (private communication). The actual details of technique and interpretation described here were evolved independently and in accordance with the particular soil circumstances and observational requirements involved.

† This and tables 3, 4 and 5 are reproduced facing pp. 128 and 129.

the mean is derived from twenty readings in observations 6 to 14, sixteen readings in observations 15 to 30. The figures are given to an accuracy consistent with the nominal accuracy of reading, though with regard to temperature and humidity it should be noted that only the differences from one height to another, and not the absolute values, are nominally accurate to 0.01°F and 0.01 g./m.^3 respectively. Since, however, the interest lies primarily in the gradients of these factors, the differences, and not the absolute values, are of paramount importance. In appraising the real accuracy of these figures the control measures adopted should be kept in mind. The first procedure, in which psychrometers and anemometers were compared at the same height, was employed in observations 6 to 14; the second and more reliable procedure of systematic interchanging being adopted for the rest of the measurements. The greatest weight is accordingly attached to observations 15 to 30, particularly as regards the differences over the intervals 200 to 100, 150 to 37.5 and 50 to 25 cm., since for these intervals it may reasonably be assumed that systematic instrumental inequalities not covered by the calibrations are automatically eliminated by the interchange process.

Interest first naturally turns to the forms of the vertical profiles, since from these we require to obtain the vertical gradients of humidity, temperature and wind speed, and furthermore we propose to utilize certain laws relating the wind profile and aerodynamic drag at the earth's surface. Earlier studies of the wind profile in particular have recognized the influence of wind speed and vertical temperature gradient on the form of the profile. On physical grounds an interdependence of these factors is to be expected, since the magnitude of the temperature gradient clearly determines the initial tendency towards modification of the flow by buoyancy forces, while the strength of the wind and the associated shearing forces constitute an opposing tendency and may be expected to control the extent of the modification. These ideas have been developed by a number of workers (see Brunt 1939, p. 242), and the general belief now is that the so-called 'stability' of the atmosphere is specified by the following number, usually termed the Richardson number,

$$\frac{g}{T} \frac{(\partial T/\partial z + \Gamma)}{(\partial u/\partial z)^2},$$

where T is in $^{\circ}\text{K}$, Γ is the dry adiabatic lapse rate, $\partial T/\partial z$ and $\partial u/\partial z$ are the vertical temperature and wind velocity gradients, z being measured upward. The sign of the above number is of course controlled by the sign of the temperature gradient, positive and negative values denoting stable and unstable thermal stratifications respectively. Zero or relatively low numerical values imply a state in which thermally induced buoyancy forces are entirely absent or dominated by large dynamical forces. For simplicity these conditions will usually be described as 'stable', 'unstable' and 'neutral'.

As an illustration of the general features of the present vertical profiles, three small groups of observations have been selected as representative of moderate instability, near-neutral conditions and moderate stability, the mean values of the Richardson number at 75 cm. being -0.056 , -0.005 and $+0.067$ respectively. The selection has been made from observations 15 to 30, which, being subject to a more

satisfactory control than the earlier observations, should provide a more critical demonstration of the forms of the profiles. The mean profiles for each property and each group are shown graphically in figure 2, a common representation of all three properties, wind speed, temperature and humidity, being achieved by plotting against the logarithm of the height, the values of the parameter

$$\frac{S_{25} - S_z}{S_{25} - S_{50}},$$

S_z representing wind speed, absolute humidity or temperature at height z cm. In neutral conditions both wind speed and absolute humidity show a close approach to a linear relation with the logarithm of the height. The remaining profiles of wind speed and humidity and the two profiles of temperature exhibit a systematic departure from linearity, such that in unstable conditions the $S/\log z$ curve is concave upwards and in stable conditions convex upwards. This is wholly consistent with recent profile results quoted by Sheppard (1947) and by Deacon (1949). The departure of individual points from straight lines or smooth curves provides some indication of the general accuracy and representative nature of the measurements. As might be expected, greatest consistency has evidently been achieved in the wind-speed and dry-bulb measurements.

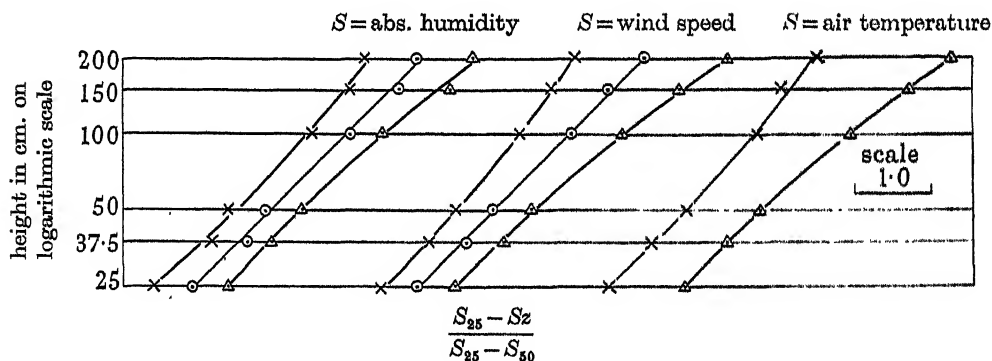


FIGURE 2. Vertical profiles of absolute humidity, wind speed and air temperature above short grass.

observation no.	mean Richardson no. at 75 cm.
× 15, 17, 20	-0.056
○ 18, 22, 26, 29	-0.005
Δ 19, 23	+0.067

Since in the following analysis considerable importance is attached to the form of the vertical wind profile, a more detailed demonstration of this feature is of some interest at this stage. Wind profiles corresponding to the individual observations in the more accurate group (nos. 15 to 30) are shown in figure 3, the parameter $\frac{u_{25} - u_z}{u_{25} - u_{50}}$ being plotted against height on a logarithmic scale. The profiles are here arranged in the order of the Richardson numbers at 75 cm., and in order to emphasize the variation of the profile with the latter number straight lines have been

drawn through the points at 25 and 50 cm. With the exception of a single observation (no. 30) these wind profiles exhibit, in a more detailed and systematic fashion, the variation with Richardson number already indicated by the three selected groups above. The essential feature is that when the Richardson number is numerically small, i.e. in near-neutral conditions, there is a linear relation between u and $\log z$. In unstable conditions (Richardson number negative) u increases less rapidly than $\log z$, while in stable conditions it increases more rapidly than $\log z$. Analytical interpretations of these profiles will be noted in subsequent stages of this paper.

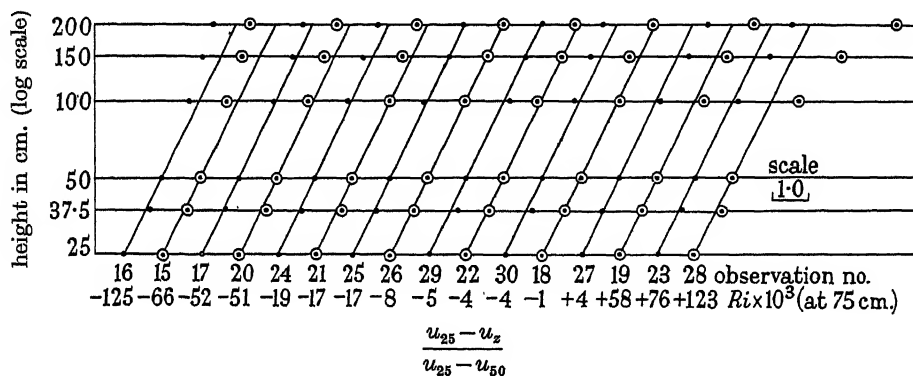


FIGURE 3. Vertical profiles of wind velocity over a short grass surface in relation to Richardson number (Ri).

In general two procedures have been followed, according to the general consistency of the profile data, in evaluating the vertical gradients of temperature, humidity and wind velocity. For observations 6 to 14 smooth curves were drawn through the individual $S/\log z$ profiles and tangents drawn to these curves at approximately the mid-point (actually 75 cm.), i.e. well away from the extremities of the curve, so that experimental scatter and personal error in drawing would be unlikely to lead to unreproducible results. For observations 15 to 30 the greater reliability attached to the differences over the various height intervals involved in the interchanging process permitted a reproducible arithmetical derivation of gradients at a number of heights. The basis of this is that except in cases of extreme curvature of the \log height curve, the variations with height of the property S may be represented approximately by the equation

$$S = a + b \log(z + c),$$

whence

$$\frac{\partial S}{\partial z} = \frac{b}{(z + c)},$$

where b and c may be derived from values of the property S at three heights. Actually c was found to vary systematically with height, so the method adopted was to obtain values of c appropriate to the lower and higher sections of the profile, using the differences 200–100, 150–37.5 and 150–37.5, 50–25 respectively. These values of c were used to deduce $\partial S/\partial z$ at 150 and 37.5 cm. respectively, while a mean of the two values was used in computing $\partial S/\partial z$ at 75 cm. The values so obtained are reproduced in table 3, u , χ and T referring to wind speed, absolute humidity and absolute temperature respectively. Values of the Richardson number are also

tabulated there. A few omissions occur in groups C , C' , D and D' when gradients could not be derived with confidence, owing either to inefficient performance of the 100 cm. wet bulb or to apparently spurious features in some part of the profile.

5. EDDY DIFFUSIVITY AND EVAPORATION IN NEUTRAL CONDITIONS

From the observed values of the rate of evaporation and vertical gradient of humidity we may now derive magnitudes of the vertical component of the eddy diffusivity for water vapour, K_v , and compare these with the corresponding magnitudes for momentum, K_m . Consideration will be restricted in the first instance to the simpler case in which the atmosphere is in a neutral state as indicated by numerically small values of the Richardson number. In specifying K_m the deficiencies of existing theoretical approaches may be avoided, following Calder (1949), by appealing directly to the well-established law for aerodynamically rough flow through pipes (Nikuradse 1933) and over flat plates (Schlichting 1936), and to the evidence for the validity of this law in the surface layers of the atmosphere. The laboratory law, written in a form convenient for meteorological application (following Prandtl 1932), is

$$u_z = \frac{1}{k} \sqrt{\left(\frac{\tau_0}{\rho}\right)} \log_e \frac{z}{z_0}, \quad (1)$$

where u_z is the mean wind speed at height z , τ_0 the horizontal shearing stress at $z = 0$, ρ the air density, k an aerodynamical constant of value 0.40, and z_0 the roughness parameter of the surface.

Differentiating equation (1) and substituting in the equation defining K_m , i.e.

$$\tau = \rho K_m \frac{\partial u}{\partial z},$$

we have, with the customary assumption (proved by Ertel 1933) of the constancy of τ with height in the lower atmosphere,

$$K_m = \frac{u_1 k^2 z}{\log_e z_1/z_0}, \quad (2)$$

where u_1 is the mean wind speed at some arbitrarily chosen level z_1 , or

$$K_m = k^2 z^2 \frac{\partial u}{\partial z}. \quad (3)$$

The applicability of the functional form of equation (1) to the lower layers of the atmosphere in neutral conditions is now widely accepted, and the profile data already discussed here provide additional confirmation of this feature for a short grass surface. However, the feature which is of most critical importance in the present analysis is the explicit validity of equation (1) in atmospheric flow. This was established recently by Sheppard (1947), who measured the drag of the earth's surface (i.e. the horizontal shearing stress) in near-neutral conditions and substituting in equation (1) obtained values for k close to 0.4. Thus, for neutral atmospheric flow near the surface, we shall assume that equations (2) and (3) with $k = 0.4$ specify true values of K_m , and proceed to compare such values with those of K_v deduced directly from the present measurements.

Making the reasonable assumption of negligible horizontal variation of absolute humidity K_v may be directly obtained from the appropriate integrated equation of transfer (neglecting variations of air density with height), as follows:

$$E + \left(K_v \frac{\partial \chi}{\partial z} \right)_z = \int_0^z \frac{\partial \chi}{\partial t} \partial z, \quad (4)$$

employing the time mean values of $\partial \chi / \partial z$, $\partial \chi / \partial t$ and rate of evaporation E . Values of E , obtained directly from the water losses shown by the 4 in. soil evaporimeters, are included in table 3. In practice, however, it was easily verified that for the shallow layer involved here the term on the right-hand side of equation (4) could be neglected. Indeed, the magnitude of

$$\frac{1}{E} \int_0^{z=200} \frac{\partial \chi}{\partial t} \partial z$$

for many of the observations was less than 0.005, while the maximum value, occurring with the low rates of evaporation of observations 9, 23 and 28, was only of the order of 0.02. Thus with sufficient accuracy

$$K_v = \frac{-E}{\partial \chi / \partial z}, \quad (5)$$

and in order to facilitate comparison with the form given for K_m in equation (3), the above expression is reduced to the non-dimensional parameter

$$\frac{K_v}{z^2 \frac{\partial u}{\partial z}} = \frac{-E}{z^2 \frac{\partial u}{\partial z} \frac{\partial \chi}{\partial z}}, \quad (6)$$

and is tabulated in this form in table 3.

If K_v and K_m are identical then from equations (3) and (6) we should have in neutral conditions

$$\sqrt{\left(\frac{-E}{z^2 \frac{\partial u}{\partial z} \frac{\partial \chi}{\partial z}} \right)} = k = 0.4.$$

Values of the above parameter, corresponding to cases with Richardson number at 75 cm. (Ri_{75}) numerically less than 0.005, have been obtained from table 3 and are reproduced below. A considerable range of wind speed is covered by these cases:

observation no.	Ri_{75}	u_{200} (cm./sec.)	height (cm.)	$\sqrt{\left(\frac{-E}{z^2 \frac{\partial u}{\partial z} \frac{\partial \chi}{\partial z}} \right)}$
8	0.000	512	75	0.39
13	0.003	244	75	0.51
18	-0.001	422	37.5	0.41
			75	0.44
			150	0.44
22	-0.004	490	37.5	0.36
			75	0.37
			150	0.39
27	0.004	376	75	0.38
			150	0.37
30	-0.004	654	75	0.46
				mean 0.41

It is seen that the above figures, though somewhat variable, as might well be expected from the nature of the measurements made, have a mean value very close to 0.40. As far as is known this provides for the first time a direct verification of the identity of the eddy diffusivity for water vapour with that for momentum.

6. THE INFLUENCE OF ATMOSPHERIC STABILITY

We may now turn to a more general consideration of the data assembled in table 3, where a substantial variation in the stability of the atmosphere is indicated by the wide range of the Richardson number, which at a height of 75 cm. amounts to -0.125 to $+0.123$. The foregoing considerations have demonstrated that in neutral conditions both wind speed and absolute humidity vary linearly with the logarithm of the height in the surface layer and that the parameter

$$\frac{-E}{z^2 \frac{\partial u}{\partial z} \frac{\partial \chi}{\partial z}} \quad \text{or} \quad \frac{K_v}{z^2 \frac{\partial u}{\partial z}}$$

is constant. In other atmospheric conditions we have already seen that the vertical profiles of wind speed and absolute humidity depart from the simple logarithmic relation in a similar and systematic fashion. It is now evident from table 3 that in such conditions the above parameter is no longer constant but varies, between very wide limits, over the range of atmospheric stability involved. An obvious step was to investigate the possibility of a connexion between this parameter and the Richardson number, and the result is shown graphically in figure 4, where all values are plotted. A consistent variation of the parameter with the Richardson number is exhibited, and there is no suggestion of a systematic separation of the points according to the height above the boundary. A close approach to a unique relation between $K_v / \left(z^2 \frac{\partial u}{\partial z} \right)$ and Richardson number is thus established for the shallow surface layer here considered. For purposes of further discussion a smooth curve has been drawn through the points and corresponding values tabulated below:

Ri	-0.25	-0.20	-0.15	-0.10	-0.05	0.00	+0.05	+0.10
$\frac{K_v}{z^2 \frac{\partial u}{\partial z}}$	0.72	0.58	0.45	0.34	0.24	0.17	0.11	0.07

An increase in the diffusive action of the atmosphere in unstable conditions, and a reduction in stable conditions, both of which are demonstrated by the above data, are to be expected on any plausible theory of turbulent transport. The main interest of the present results lies partly in the magnitude of the effect, and we note that a tenfold variation in the parameter $K_v / \left(z^2 \frac{\partial u}{\partial z} \right)$ occurs over the range of stability covered here. However, we should also note that the high numerical values of the Richardson number, and hence the greatest modification of the above parameter, are associated with the 150 cm. level. The numerical values of the Richardson number decrease rapidly as the ground is approached, and at the lowest level the range of numbers computed here is only -0.035 to $+0.045$. The corresponding

TABLE 2. WIND SPEED, AIR TEMPERATURE AND HUMIDITY DATA

group and date	A, 10. iii. 48										B, 18. iii. 48					C, 24. iii. 48					C', 25. iii. 48					D, 28. iii. 48					D', 29. iii. 48				
obs. no.	...	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30									
G.M.T.*	...	10.33	12.30	15.30	17.30	19.30	12.35	14.30	16.34	19.32	09.38	11.30	14.30	16.30	18.30	10.30	12.35	15.30	18.30	10.30	12.33	14.42	16.30	19.00	11.39	14.03									
wind speed in cm./sec. at heights in cm.	...	443	379	512	282	357	386	280	244	241	253	268	323	422	277	327	479	490	250	522	570	525	376	197	781	654									
200	443	379	512	282	357	386	386	280	244	241	253	268	323	422	277	327	479	490	250	522	570	525	376	197	781	654									
150	426	365	484	266	337	370	370	268	232	228	247	249	313	398	266	314	458	469	291	500	547	500	376	166	743	624									
100	403	361	468	247	319	354	354	257	220	205	233	229	298	382	238	298	435	435	204	472	506	484	383	142	707	589									
50	358	308	401	207	272	309	309	224	190	172	210	215	265	336	200	264	381	385	108	420	450	407	293	106	612	514									
37.5	385	292	+	191	257	257	293	211	181	163	200	206	253	317	189	254	365	363	157	398	427	386	276	101	594	493									
25	299	257	335	163	223	223	268	191	161	143	178	184	231	292	170	225	325	330	136	360	389	349	263	85	594	426									
air temp. in ° F at heights in cm.	200	51.7	54.2	56.2	53.1	48.8	48.9	49.9	50.6	44.4	47.9	52.2	55.1	53.0	45.0	54.5	53.3	58.8	51.0	54.1	56.9	57.6	56.4	47.8	56.7	56.4									
150	51.88	54.39	56.20	52.80	48.60	49.16	49.16	50.11	50.47	44.16	48.11	52.45	55.29	53.01	44.71	54.80	58.47	58.89	50.52	54.50	57.13	57.71	56.39	47.18	56.95	56.52									
100	52.15	54.69	56.20	52.34	48.32	49.44	49.44	50.18	50.34	43.79	48.21	52.70	55.55	53.00	44.30	54.93	58.66	58.94	49.85	54.74	57.54	57.86	56.35	46.39	57.03	56.64									
50	52.72	55.35	56.18	51.41	47.80	50.02	50.02	50.49	50.33	43.19	48.68	53.38	55.96	53.01	43.66	55.45	59.04	59.07	48.82	55.27	58.15	58.11	56.18	45.19	57.40	56.68									
37.5	53.00	55.66	56.19	51.05	47.61	50.32	50.32	50.66	50.32	42.97	48.81	53.58	56.22	53.03	43.43	55.74	59.22	59.12	48.44	55.57	58.41	58.23	56.26	44.85	57.60	57.60									
25	53.39	56.06	56.16	50.55	47.34	50.68	50.68	50.86	50.25	42.65	49.09	53.92	56.45	53.00	43.14	56.07	59.43	59.16	47.91	55.77	58.76	58.37	56.28	44.31	57.65	56.74									
abs. humidity in g./m. ³ at heights in cm.	200	6.07	5.97	6.04	5.84	7.68	5.34	4.99	4.78	5.21	4.67	5.36	5.13	5.01	4.47	4.95	4.34	3.26	3.70	4.94	+	5.36	5.60	5.39	8.11	7.63									
150	6.13	6.07	6.11	5.90	7.69	7.69	5.48	5.07	4.84	5.23	4.74	5.46	5.19	5.06	4.48	4.99	4.34	3.30	3.74	4.90	+	5.42	5.65	5.40	8.12	7.61									
100	6.24	6.17	6.17	5.93	7.75	7.75	5.68	5.19	4.97	5.27	4.86	5.51	5.31	5.17	4.56	5.13	4.46	3.46	3.86	5.02	5.31	5.53	5.76	5.45	8.21	7.68									
50	6.41	6.45	6.39	6.09	7.79	7.79	5.89	5.46	5.13	5.35	5.13	5.85	5.63	5.37	4.63	5.45	4.75	3.74	4.01	5.15	5.39	5.74	5.98	5.65	8.38	7.75									
37.5	6.49	6.50	6.40	6.13	7.84	7.84	6.04	5.54	5.19	5.37	5.21	5.95	5.70	5.40	4.68	5.49	4.82	3.78	4.05	5.25	5.60	5.81	5.95	5.53	8.41	7.81									
25	6.61	6.68	6.49	6.18	7.86	7.86	6.17	5.67	5.29	5.42	5.39	6.18	5.95	5.55	4.73	5.69	5.03	3.97	4.13	5.39	5.70	5.93	6.13	5.68	8.51	7.89									

* Refers to midtime of observation period of approximately 60 min.

† Anemometer reading obviously spurious.

‡ Wet-bulb dry at tip.

TABLE 3. REDUCED DATA FOR MOISTURE-EXCHANGE ANALYSIS

group and date		A, 10. iii. 48										B, 18. iii. 48										C, 24. iii. 48										C, 26. iii. 48										D, 28. iii. 48										D, 29. iii. 48			
obs. no.	...	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30																													
G.M.T.*	...	10.38	12.30	15.30	17.30	19.30	12.35	14.30	16.34	18.32	09.38	11.30	14.30	16.30	18.30	10.30	12.35	15.30	18.30	10.30	12.33	14.42	16.30	19.00	11.39	14.03																													
U at 2 m. (cm./sec.)	...	443	379	512	282	357	386	280	244	241	263	258	323	422	277	327	479	430	250	522	570	525	376	197	781	654																													
wind direction	...	WNW	NW	WNW	WNW	WNW	NNW	NNW	W	SSW	SW	S	SSE	SSE	SSE	ESE	ESE	ESE	ESE	ESE	ESE	ESE	SSE	SE	S	SSW																													
cloud	...	3/10 Ci	3/10 Ci	3/10 CiCs	3/10 Ci	1/10 Ci	5/10 Cu	5/10 Cu	2/10 Cu	8/10 CiCs	nil	nil	tr. Ci	1/10 Ci	2/10 Ci	3/10 Ci	7/10 Ci	3/10 Ci	7/10 Ci	tr. Ac	tr. Ac	tr. Cu	2/10 Ac	1/10 Ac	6/10 Ac	8/10 CuS																													
$10^8 \times E/(G./cm.^2 \text{ sec.})$		340	355	275	80	65	610	395	235	85	345	410	390	260	75	320	420	350	75	430	440	330	185	25	420	255																													
$-10^{10} \times \frac{\partial \gamma}{\partial z} (G./cm.^4)$		37.5 cm.	42	29	22	11	53	46	33	14	97	119	117	69	35	94	103	91	47	85	108	75	—	—	50	49																													
$10^8 \times \frac{\partial u}{\partial z} (\text{sec.}^{-1})$		75 cm.	33								44	44	47	32	19	46	44	46	30	31	37	37	57	13	28	17																													
$10^8 \times \frac{\partial u}{\partial z} (\text{sec.}^{-1})$		150 cm.									18	14	17	15	9	17	—	19	16	—	—	16	29	—	10	—																													
$10^8 \times \frac{\partial u}{\partial z} (\text{sec.}^{-1})$		37.5 cm.									115	109	128	165	117	142	208	209	125	224	234	223	137	—	313	317																													
$10^4 \times \frac{\partial T}{\partial z} (^{\circ} K/cm.)$		75 cm.	85	67	109	71	73	53	49	63	44	40	56	73	65	57	87	99	72	97	116	110	79	60	143	125																													
$10^4 \times \frac{\partial T}{\partial z} (^{\circ} K/cm.)$		150 cm.									19	19	24	39	38	28	39	53	45	48	62	59	41	—	71	63																													
$10^8 E$		37.5 cm.									—85	—116	—104	+4	+114	—125	—83	—20	+198	—107	—131	—55	—	—	—54	—																													
$-\frac{\partial \chi \partial u}{\partial z \partial z} z^2$		75 cm.	—58	—66	+1	+93	—61	—34	+1	+64	—37	—60	—49	—2	+69	—49	—40	—12	+113	—56	—68	—28	+7	+125	—30	—19																													
$10^8 E$		150 cm.									—16	—27	—24	0	+38	—23	—	—8	+62	—	—	—14	+3	—	—18	—																													
$-\frac{\partial \chi \partial u}{\partial z \partial z} z^2$		37.5 cm.									220	225	185	165	125	170	140	130	90	*160	125	140	—	—	190	—																													
$10^8 E$		75 cm.	215	225	155	90	280	290	260	175	320	410	260	195	105	215	195	135	60	255	180	145	145	60	190	210																													
$10^8 E$		150 cm.									445	690	435	195	100	295	—	150	45	—	—	150	135	—	265	—																													
$10^8 g \left(\frac{\partial T}{\partial z} + 1 \right)$		37.5 cm.									—22	—38	—22	+1	+29	—21	—7	—1	+44	—7	—9	—4	—	—	—2	—																													
$T \left(\frac{\partial u}{\partial z} \right)^2$		75 cm.	—27	—50	0	+64	—39	—40	+3	+57	—66	—125	—52	—1	+58	—51	—17	—4	+76	—19	—17	—8	+4	+123	—5	—																													
$T \left(\frac{\partial u}{\partial z} \right)^2$		150 cm.									—144	—245	—140	+2	+95	—97	—	—8	+107	—	—	—12	+8	—	—11	—																													

* Refers to midtime of observation period of approximately 60 min.

TABLE 4. SOIL TEMPERATURE DATA AND HEAT CAPACITIES

group and date	A, 10. iii. 48					C, 24. iii. 48					C', 25. iii. 48					D, 28. iii. 48				
obs. no.	6	7	8	9	10	15	16	17	18	19	20	24	25	28	20	24	25	28		
G.M.T.*	10.33	12.30	15.30	17.30	19.30	09.38	11.30	14.30	16.30	18.30	10.30	10.30	12.33	19.00	10.30	10.30	12.33	19.00		
soil temperature in ° F at depths in in.	0	52.25	57.05	49.55	46.25	52.90	61.85	61.30	53.35	43.25	57.80	55.35	61.95	44.40	57.80	55.35	61.95	44.40		
	1	49.65	53.35	54.70	50.30	47.50	53.60	56.30	52.50	46.35	49.45	49.35	55.20	47.05	49.45	49.35	55.20	47.05		
	2	48.45	51.75	52.80	50.60	48.10	44.35	49.40	52.25	48.25	45.75	46.30	51.70	48.65	45.75	46.30	51.70	48.65		
	4	47.40	50.00	51.90	50.70	48.75	42.85	46.65	51.85	49.40	43.20	44.25	48.80	49.45	43.20	44.25	48.80	49.45		
	6	46.55	47.95	50.05	50.00	49.15	42.55	44.25	48.55	50.00	49.60	41.80	42.85	45.60	41.80	42.85	45.60	49.30		
	8	46.30	46.80	48.35	48.85	48.70	43.35	43.80	46.60	48.00	48.50	42.30	43.05	44.30	42.30	43.05	44.30	48.15		
	12	46.20	46.20	47.05	47.55	47.85	44.35	44.10	45.50	46.50	47.30	43.20	43.80	44.05	43.20	43.80	44.05	46.05		
	16	45.95	45.80	46.85	46.25	46.55	45.75	45.35	45.45	45.70	46.15	44.95	45.25	44.90	44.95	45.25	44.90	45.60		
	final temp. - initial temp. in ° F (effective period 57 min.) at depths in in.	45.25	45.20	45.20	45.35	45.60	46.15	45.85	45.85	45.80	45.85	45.65	45.75	45.45	45.45	45.65	45.75	45.45	45.45	
	0	+2.80	+1.70	-1.45	-4.30	-1.90	+6.05	+2.65	-2.15	-6.40	-4.15	+2.00	+4.95	+0.95	-1.10	+2.00	+4.95	+0.95	-1.10	
	1	+1.80	+1.60	-1.05	-2.40	-1.25	+3.75	+2.40	-0.60	-3.65	-2.50	+2.60	+3.45	+1.80	-1.65	+2.60	+3.45	+1.80	-1.65	
	2	+1.55	+1.50	-0.65	-1.60	-1.25	+2.50	+2.25	+0.25	-1.90	-1.75	+2.60	+2.80	+2.10	-1.65	+2.60	+2.80	+2.10	-1.65	
	4	+0.90	+1.20	-0.30	-0.85	-1.10	+1.50	+2.05	+0.80	-0.70	-1.35	+2.35	+2.10	+2.00	-1.25	+2.35	+2.10	+2.00	-1.25	
	6	+0.25	+0.80	+0.20	-0.20	-0.75	+0.40	+1.30	+1.00	+0.20	-0.55	+1.10	+0.90	+1.50	-0.60	+1.10	+0.90	+1.50	-0.60	
	8	-0.15	+0.35	+0.30	+0.10	-0.55	-0.15	+0.45	+0.80	+0.40	0.00	+0.40	+0.20	+0.80	+0.15	+0.40	+0.20	+0.80	+0.15	
	12	-0.35	+0.10	+0.20	+0.25	-0.30	-0.30	+0.05	+0.50	+0.45	+0.35	-0.15	-0.10	+0.35	+0.30	-0.15	-0.10	+0.35	+0.30	
16	-0.30	-0.10	+0.10	+0.10	-0.25	-0.25	-0.30	0.00	+0.05	+0.05	-0.30	-0.25	-0.05	+0.35	-0.30	-0.25	-0.05	+0.35		
date of soil sampling	-0.15	-0.10	-0.00	+0.05	-0.25	-0.05	-0.30	-0.10	-0.05	-0.05	-0.10	-0.15	-0.05	+0.20	-0.10	-0.15	-0.05	+0.20		
layer (in.)	10. iii. 48					24. iii. 48					29. iii. 48									
apparent sp.gr.	0-1	1-2	2-4	4-6	6-8	0-1	1-2	2-4	4-6	6-8	0-1	1-2	2-4	4-6	0-1	1-2	2-4	4-6		
% by wt. of water	1.38	1.69	1.71	1.76	1.79	1.80	1.85	1.82	1.61	1.77	1.80	1.85	1.82	1.61	1.31	1.55	1.66	1.68		
apparent sp.ht.	23.3	19.7	20.6	21.3	20.5	19.8	17.8	18.6	19.5	19.9	24.3	17.0	18.2	19.4	24.3	17.0	18.2	20.1		
heat capacity per unit volume (cal/° C cm. ³)	0.39	0.36	0.37	0.37	0.36	0.36	0.34	0.35	0.36	0.36	0.39	0.34	0.34	0.35	0.39	0.34	0.34	0.36		
	0.53	0.61	0.62	0.65	0.65	0.64	0.63	0.63	0.57	0.64	0.52	0.52	0.57	0.60	0.52	0.52	0.57	0.61		

* Refers to midtime of observation period of approximately 60 min.

TABLE 5. HEAT-BALANCE ANALYSIS

group and date	A, 10. iii. 48										C, 24. iii. 48					C', 25. iii. 48				D, 28. iii. 48			
	6	7	8	9	10	15	16	17	18	19	15	16	17	18	19	20	24	25	28				
...	10.33	12.30	15.30	17.30	19.30	09.38	11.30	14.30	16.30	18.30	09.38	11.30	14.30	16.30	18.30	10.30	10.30	12.33	19.00				
...	16.3	17.9	13.4	7.3	5.7	14.4	18.0	16.3	11.2	3.8	14.4	18.0	16.3	11.2	3.8	17.6	15.0	18.3	3.4				
...	340	355	275	80	65	345	410	390	260	75	345	410	390	260	75	320	430	440	25				
...	0.715	0.770	0.336	0.031	nil	0.637	0.835	0.692	0.314	0.006	0.637	0.835	0.692	0.314	0.006	0.764	0.803	0.903	nil				
...	0.123	0.128	0.073	0.008	nil	0.129	0.157	0.137	0.071	0.003	0.129	0.157	0.137	0.071	0.003	0.137	0.156	0.168	nil				
...	0.191	0.196	0.153	0.116	0.109	0.186	0.200	0.185	0.155	0.119	0.186	0.200	0.185	0.155	0.119	0.195	0.180	0.192	0.099				
...	0.057	0.105	0.017	-0.048	-0.063	0.117	0.154	0.089	-0.042	-0.088	0.117	0.154	0.089	-0.042	-0.088	0.135	0.129	0.148	-0.053				
...	0.120	0.126	0.097	0.029	0.023	0.121	0.144	0.138	0.092	0.026	0.121	0.144	0.138	0.092	0.026	0.112	0.152	0.155	0.009				
...	0.224	0.215	-0.004	-0.074	-0.069	0.084	0.180	0.143	0.038	-0.054	0.084	0.180	0.143	0.038	-0.054	0.185	0.186	0.240	-0.055				
...	-57	-65	+2	+94	+54	-36	-59	-48	-1	+70	-36	-59	-48	-1	+70	-48	-55	-67	+126				
...	2150	1820	—	430	690	1270	1680	1650	—	420	1270	1680	1650	—	420	2140	1890	2020	240				
...	1030	840	950	370	570	780	930	830	810	390	780	930	830	810	390	690	1400	1180	190				
...	450	485	—	105	165	520	740	520	—	115	520	740	520	—	115	670	345	310	70				
...	215	225	155	90	135	320	410	260	195	105	320	410	260	195	105	215	255	180	60				
...	-27	-50	0	+64	+33	-66	-125	-52	-1	+58	-66	-125	-52	-1	+58	-51	-19	-17	+123				

smoothed values of the parameter $K_v / \left(z^2 \frac{\partial u}{\partial z} \right)$ are 0.23 and 0.11, i.e. a variation of about 30 % on the value of 0.17 at zero Richardson number. Thus, at very low heights, say below 50 cm., the influence of a moderate range of stability could be neglected, and equations (1) and (2) applied to give the eddy diffusivity, without incurring errors greater than about 30 %. At height ranges more customarily adopted, however, the error so introduced would be very large, and we therefore require to give further consideration to the present results with particular reference to the effect of atmospheric stability.

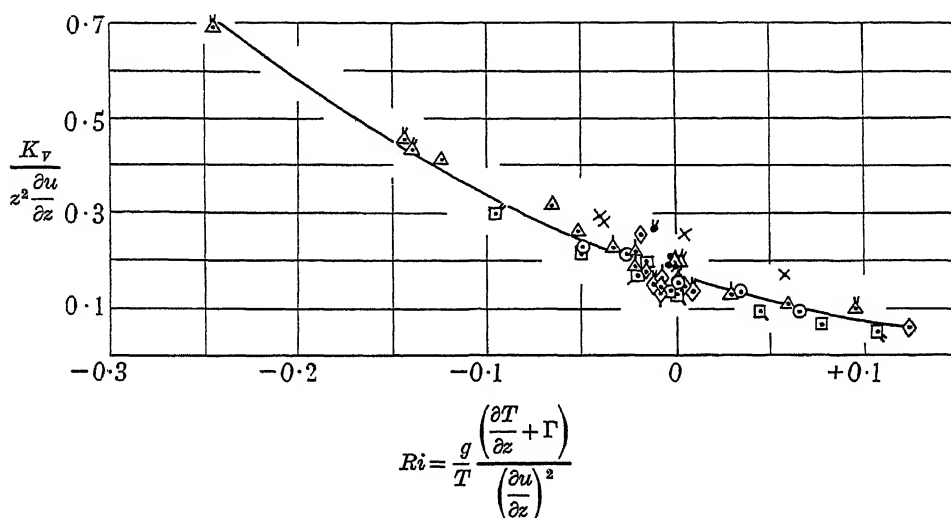


FIGURE 4. Relation between eddy diffusivity for water vapour (K_v), vertical wind gradient ($\partial u / \partial z$) and Richardson number (Ri) at various heights above a short grass surface.

group	symbol	
A	⊙	
B	×	
C	△	undashed points $z = 75$
C'	◻	single dash $z = 37.5$
D	◊	double dash $z = 150$
D'	●	

An attempt to prescribe theoretically the effects of atmospheric stability on the form of the wind profile in the boundary layer has been made by Rossby & Montgomery (1935). This treatment, the details of which it is not proposed to discuss in this paper, leads to an expression relating wind shear and horizontal shearing stress, from which the following equation may be derived:

$$\frac{K_m}{z^2 \frac{\partial u}{\partial z}} = \frac{k^2}{(1 + \sigma Ri)}, \quad (7)$$

where k is von Karman's constant and σ is a proportionality factor introduced in the relation between eddying energy and potential energy associated with thermal stratification. Deacon (1949) finds that the Rossby-Montgomery formula does not

agree with wind-gradient observations at large stability and instability, and demonstrates that a modified formula, proposed by Holzman (1943) entirely on empirical grounds and in the interests of mathematical tractability, provides a better representation. Holzman's formula may be reduced to the form

$$\frac{K_m}{z^2 \frac{\partial u}{\partial z}} = k^2(1 - \sigma Ri), \quad (8)$$

which it will be seen is identical with equation (7) when the Richardson number is small.

Equations (7) and (8) are readily compared with the experimental results summarized above, and it is found that neither equation can satisfactorily describe the whole range of results. It is noteworthy, however, that reasonable agreement with the results in stable conditions is provided by the Rossby-Montgomery equation with σ equal to about 12, while in unstable conditions the Holzman formula gives corresponding agreement using the same value of σ . No physical explanation is immediately evident for this feature, nor for the magnitude attached to σ . Thus, in both functional and explicit form the equations put forward by Rossby & Montgomery and Holzman fail to provide a completely satisfactory basis for expressing the observed relation between eddy diffusivity and Richardson number. To what extent this is due to deficiencies in the derivation of these equations or to a genuine difference between K_p and K_m in non-neutral conditions is somewhat problematical.

An alternative basis for examining the present results is to be found in the generalized wind-profile law recently put forward by Deacon (1949), who found that an extensive series of wind profiles, exhibiting characteristics similar to those shown in figure 3, was fairly well represented by the equation

$$\frac{\partial u}{\partial z} = az^{-\beta}, \quad (9)$$

in which the exponent β is > 1 , $= 1$ and < 1 in unstable, neutral and stable conditions respectively. Integrating this empirical law and making use of the observed decrease in stability influence on the wind profile as the boundary is approached, Deacon obtained

$$u_z = \frac{1}{k(1-\beta)} \sqrt{\left(\frac{\tau}{\rho}\right) \left\{ \left(\frac{z}{z_0}\right)^{1-\beta} - 1 \right\}}, \quad (10)$$

which reduces to equation (1) for small values of z and values of β not differing greatly from unity. Actually it was found that values of z_0 evaluated from equation (10) showed a considerable variation with stability, which is inconsistent with the observed decrease of stability influence on the wind profile as the surface is approached. Deacon suggests that the variation is probably due to β not being quite constant with height, except possibly in very stable conditions when aerodynamically smooth flow may conceivably occur near the surface and so lead to a genuine reduction in the roughness parameter. In order to avoid this inconsistency in the application of equation (10) Deacon makes the assumption that z_0 is independent of stability (except possibly in very stable conditions), its value being prescribed by equation (1) and wind-profile

measurements in neutral conditions. Using this value of z_0 and wind-profile measurements over the same surface in other conditions of stability equation (10) is then employed to calculate a mean value of β for the height layer involved.

From equation (10) it is easily shown that K_m is given by

$$K_m = k \sqrt{\left(\frac{\tau}{\rho}\right) z_0^{1-\beta} z^\beta} = k^2 z^2 \frac{\partial u}{\partial z} \left(\frac{z}{z_0}\right)^{2\beta-2},$$

$$\text{or} \quad \frac{K_m}{z^2 \frac{\partial u}{\partial z}} = k^2 \left(\frac{z}{z_0}\right)^{2\beta-2}, \quad (11)$$

a form which is easily comparable with our results for K_r and requires only the derivation of z_0 and β , for which the following process was adopted. For observations nos. 15 to 30, values of the wind ratio $u_{150}/u_{37.5}$ (which involves the largest height interval to which instrument interchanging was applied) were plotted against the Richardson number at 75 cm. and found to lie on a fairly smooth curve. The interpolated value of this ratio for $Ri = 0$ was 1.275, which with equation (1) gave $z_0 = 0.25$ cm. From equation (10)

$$\frac{u_{150}}{u_{37.5}} = \frac{150^{1-\beta} - z_0^{1-\beta}}{37.5^{1-\beta} - z_0^{1-\beta}}, \quad (12)$$

and using the above value of z_0 expression (12) was evaluated for a range of values of β . A graphical relation was thus obtained between β and the Richardson number and hence between the latter and the parameter $k^2 \left(\frac{z}{z_0}\right)^{2\beta-2}$, with $k = 0.4$ and $z = 75$ cm.

Interpolated values of the various parameters, together with the observed magnitudes of $K_r / \left(z^2 \frac{\partial u}{\partial z}\right)$, are tabulated below:

Richardson no. at 75 cm.	-0.125	-0.10	-0.05	+0.05	+0.10	+0.125
$u_{150}/u_{37.5}$	1.209	1.216	1.237	1.370	1.525	1.640
β	1.085	1.075	1.045	0.915	0.790	0.715
$k^2 \left(\frac{z}{z_0}\right)^{2\beta-2} = \frac{K_m}{z^2 \frac{\partial u}{\partial z}}$	0.42	0.37	0.26	0.06	0.015	0.005
$-\frac{E}{z^2 \frac{\partial \chi}{\partial z} \frac{\partial u}{\partial z}} = \frac{K_r}{z^2 \frac{\partial u}{\partial z}}$	0.39	0.34	0.24	0.11	0.07	0.06

($z = 75$ cm.)

A most striking agreement between the computed $K_m / \left(z^2 \frac{\partial u}{\partial z}\right)$ and the observed $K_r / \left(z^2 \frac{\partial u}{\partial z}\right)$ occurs in unstable conditions, the difference between the two values being less than 10 %, which is almost as good as the agreement obtained in the corresponding comparison in neutral conditions. With the increase of atmospheric

F. Pasquill

stability, a considerable disparity appears in the two sets of values, which is consistent with the previous suggestion that equation (10), with z_0 independent of stability, is invalid at marked stability. It is important to realize here that equation (10), and hence equation (11), though fairly well established in functional form, has not been verified explicitly. Thus, the present considerations differ from those previously applied to neutral conditions in that a direct test of the identity of K_v and K_m cannot be made. However, these considerations do show that for the range of unstable conditions involved

$$K_v(\text{observed}) = K_m(\text{computed from equations (10) and (11)}),$$

provided, of course, the roughness parameter, z_0 , is determined from equation (1) and wind-profile measurements in neutral conditions over the same surface.

7. EVALUATION OF THE HEAT-EXCHANGE COMPONENTS

The balance of heat-exchange processes near the earth's surface may be expressed as follows. If

R_s is the incoming solar radiation (direct and diffuse),

R_r the reflected component of solar radiation,

R_z the upward flux of long-wave radiation (i.e. the net effect of radiation from ground and atmosphere), at height z ,

S the heat absorbed in soil,

L the heat associated with evaporation from the ground, and

Q_z the upward turbulent flux of heat at height z ,

then, in the absence of advection effects, and neglecting the minute absorption of solar radiation in photosynthetic processes,

$$R_s = R_r + S + L + R_z + Q_z + \int_0^z c_p \rho \frac{\partial T}{\partial t} dz, \quad (13)$$

T , ρ and c_p being the absolute temperature, density and specific heat at constant pressure of the atmosphere. From the present measurements all terms with the exception of Q_z may be evaluated. Equation (13) may thus be solved for Q_z , and finally, since by definition

$$Q_z = -\rho c_p K_H \left(\frac{\partial T}{\partial z} + \Gamma \right), \quad (14)$$

the magnitude of the eddy diffusivity for heat, K_H , may be derived. Evaluation of the terms R_s , R_r and L follows without further explanation. The term

$$\int_0^z c_p \rho \frac{\partial T}{\partial t} dz,$$

which represents the rate at which sensible heat is absorbed by the air layer up to height z , was invariably less than 0.001 cal./cm.²min. and will be omitted from further considerations. The terms S and R_z remain to be discussed.

The rate at which heat is absorbed in the soil per unit area of ground surface is given by

$$S = \int_0^{\infty} \rho'c \frac{\partial \theta}{\partial t} dz,$$

or

$$\int_0^z \rho'c \frac{\partial \theta}{\partial t} dz - \left(k' \frac{\partial \theta}{\partial z} \right)_z, \quad (15)$$

where z is the depth below the soil surface, θ , ρ' , c , k' are the temperature, density, apparent specific heat and thermal conductivity of the soil medium. If, as frequently occurred in the present measurements, the soil temperature θ passes through a maximum or minimum value at depth z' then

$$S = \int_0^{z'} \rho'c \frac{\partial \theta}{\partial t} dz, \quad (16)$$

which may be readily evaluated graphically from the soil-temperature measurements, using time mean values of $\partial \theta / \partial t$ and determined values for $\rho'c$ at various depths. The apparent density of the soil, ρ' , is obtainable directly from the initial weighings of the soil samples. Subsequent drying and reweighing gives the proportion of soil material and water, and from this and the specific heat of the soil material the apparent specific heat of the soil medium may be derived by simple proportion, the air component being neglected. Determination of the specific heat of the soil material had already been carried out by the School of Agriculture, on oven-dried samples taken in the previous year from the present site, from depths 0 to 1, 2 to 3, 4 to 5 and 6 to 7 in. below the surface. The method of mixtures was employed, errors due to heat of wetting of the material being avoided by mixing it with a known quantity of water and performing the determination on the mixture. No systematic variation with depth was apparent in the values of specific heat so obtained, and the mean result was 0.20 ± 0.02 . The value 0.20 was employed in all subsequent calculations.

The soil-temperature data are summarized in table 4 in the form of mean soil temperatures at the various depths for each of the observations which included radiation measurements. With these are given the corresponding changes of temperature during the observation period, from which mean values of $\partial \theta / \partial t$ may be derived by dividing by the effective observation period (57 min.). Since the change of soil temperature occurs slowly and smoothly little error is likely to result from this simplified procedure. At the bottom of table 4 are given the data relevant to the three sets of soil samples, the values of $\rho'c$ finally obtained being used in conjunction with the corresponding temperature observations. It may be noted that the water content of the 0 to 1 in. layer includes that associated with the vegetation cover and in the evaluation of $\rho'c$ for this layer it is implicitly assumed that the specific heat of the solid material of the vegetation is the same as that of the soil solids. Since the vegetation cover could only represent a minute fraction of the mass of the 0 to 1 in. layer it is obvious that the error so involved is completely negligible. From graphs of $\rho'c$ against the mean depth of layer, smooth values were read off for the levels of temperature measurement (down to 8 in.). For observations in which z' was less than 8 in. (nos. 9, 10, 15, 16, 19, 20, 24, 25 and 28) values of $\rho'c \frac{\partial \theta}{\partial t}$ were

then plotted against depth z and expression (16) evaluated by measuring the area between the z -axis and the curve so obtained, with a planimeter, over the limits $z = 0, z = z'$.

In the cases when the soil temperature did not pass through a maximum or minimum above the 8 in. level it was necessary to use equation (15) with $z = 8$ in., and to evaluate k' at this level. This was readily achieved from the previous cases (z' less than 8 in.), since we may write

$$\int_{z'}^{z=8 \text{ in.}} \rho' c \frac{\partial \theta}{\partial t} dz = \left(k' \frac{\partial \theta}{\partial z} \right)_{z=8 \text{ in.}}, \quad (17)$$

the left-hand side of the equation being evaluated as above and $\partial \theta / \partial z$ being obtained by drawing tangents to the soil-temperature profiles. Inspection of the soil-temperature data summarized in table 4 indicates the small value of $\partial \theta / \partial z$ normally occurring in the region of 8 in. depth, so that in general the term $k' \frac{\partial \theta}{\partial z}$ is of the nature of a small correction. The magnitudes of k' obtained as described from four observations (nos. 9, 10, 19 and 20) in which the magnitudes of $\partial \theta / \partial t$ and $\partial \theta / \partial z$ were highest, were 20, 58, 18 and 35×10^{-4} cal./°C cm. sec. respectively, and a mean value of 33×10^{-4} was used throughout for the few observations (nos. 6, 7, 8, 17 and 18) where it was necessary to apply equation (15).

The long-wave radiation component from ground and atmosphere was derived by the empirical method described by Robinson (1947), employing for convenience the radiation chart described on p. 146 of that article so as to obtain directly the net outward radiation at a given level. The 'water-path'-temperature curve below 2 m. was constructed from the present temperature and humidity profiles, and the 'radiative temperature' of the ground measured as previously described, using intervals 0 to 37.5, 37.5 to 50, 50 to 75, 75 to 100, 100 to 150, 150 to 200 cm. For the remainder of the curve the necessary values of upper-air temperature and humidity (starting from the 1000 or 950 mb. levels, according as the surface pressure was above or below 1025 mb.) were interpolated from previous and succeeding Downham Market ascents. In all cases inspection of the synoptic situation suggested that no substantial differences in air-mass properties were likely to exist over the distances separating the present site and Downham Market (30 miles). Furthermore, the observations for which these computations were made were carried out in good visibility with in general less than 3/10 diffuse high cloud (see cloud data in table 3), the cloud being in most cases at a fairly low angular elevation. No attempt has been made to allow for the effects of these small cloud amounts. The worst case is observation no. 20, when the mean cloud amount was 3/10 Ci and 2/10 Ac, and this has been included deliberately in the present analysis so as to provide some appraisal of the errors likely to accrue from neglect of the cloud factor.

8. MAGNITUDES OF THE EDDY DIFFUSIVITY FOR HEAT

The measured and computed heat components are listed in table 5, and the balance is attributed to turbulent transport. The values are specified to 0.001 cal./cm.² min., though we shall now see that errors considerably in excess of this may

be expected to occur. For the estimation of long-wave radiation, Robinson (1947) suggests that the net outward radiation with clear skies is given by the radiation chart with a probable error of $\pm 0.015 \text{ cal./cm.}^2\text{min.}$ The comparison of the solarimeter with a Gorczynski or Ångström instrument suggested that the indications of the former may be in error up to $\pm 2 \%$, i.e. $\pm 0.015 \text{ cal./cm.}^2\text{min.}$ in the maximum values of incoming solar radiation measured in the present investigation. As for the errors likely to result from heterogeneity of the thermal properties of the soil it may be noted that if point-to-point variations in soil-moisture content are of the order of the differences shown by the separate measurements summarized at the bottom of table 4, then these could lead to errors of up to $\pm 10 \%$ in the heat capacity per unit volume of soil and hence $\pm 0.015 \text{ cal./cm.}^2\text{min.}$ in the maximum value of the heat absorbed in the soil. These few examples suffice to indicate the very considerable casual error which might accumulate in the balance of the various factors. Furthermore, the present observations include two cases in which a direct assessment of the resultant error is possible, since for observations nos. 8 and 18, which were carried out on widely separated occasions, the gradient of potential temperature was practically zero. In point of fact the values at 75 cm. were respectively $2 \times 10^{-4} \text{ C/cm.}$ and $-1 \times 10^{-4} \text{ C/cm.}$ respectively. The turbulent transport of heat should therefore be very small and the remaining components, i.e. those measured and estimated, should almost balance. Actually it will be seen from table 5 that the greater discrepancy occurs in no. 18 and there amounts to $0.04 \text{ cal./cm.}^2\text{min.}$, a figure which seems fairly consistent with the individual errors cited above.

We now apply equation (14) and derive apparent values of K_H for the remaining observations. The appropriate magnitudes of $\left(\frac{\partial T}{\partial z} + \Gamma\right)$ at 75 cm. and the resulting values of K_H at this level are reproduced in table 5, together with the corresponding values of K_V . It will be seen that the two diffusivities are of similar orders of magnitudes and range from 10^2 to 10^3 . A more critical comparison is conveniently made in terms of the non-dimensional parameters

$$K_H \left/ \left(z^2 \frac{\partial u}{\partial z} \right) \right. \quad \text{and} \quad K_V \left/ \left(z^2 \frac{\partial u}{\partial z} \right) \right.,$$

which are also tabulated in table 5 and plotted against the Richardson number at the same level in figure 5. A quite remarkable identity is shown between K_H and K_V in stable conditions, while with the increase of instability the former coefficient is systematically the greater of the two. With the exception of observation no. 20 all values of K_H conform to the line drawn through them to a degree which corresponds to errors in the turbulent transport component of well within $\pm 0.04 \text{ cal./cm.}^2\text{min.}$ This consistency is better than might be anticipated from the previous discussion of the magnitudes of casual errors.

A feature which we have not considered so far is the possibility of systematic error in unstable conditions. In order so to explain the observed differences in K_H and K_V the error would have to be in the form of an *overestimation of the turbulent heat flux* to an extent of about 100 %, or a substantial underestimation in the heat of evaporation term. Considering the heat-balance terms there is no reason to suspect

any appropriately systematic error in the measured terms, solar radiation, soil heat and heat of evaporation, and the only other term is the computed long-wave radiation flux. In the computation of this factor the effects of small amounts of high cloud were neglected, but this would result in a slight overestimation of the upward flux of long-wave radiation and hence an *underestimation* of the turbulent heat flux. Furthermore, the computation assumes the radiative temperature of the ground to be given in all cases by a spirit thermometer laid on the ground. Actually Robinson's unpublished data suggest that with the sun's altitude above 30° the method overestimates the radiative temperature. As far as can be inferred from this somewhat

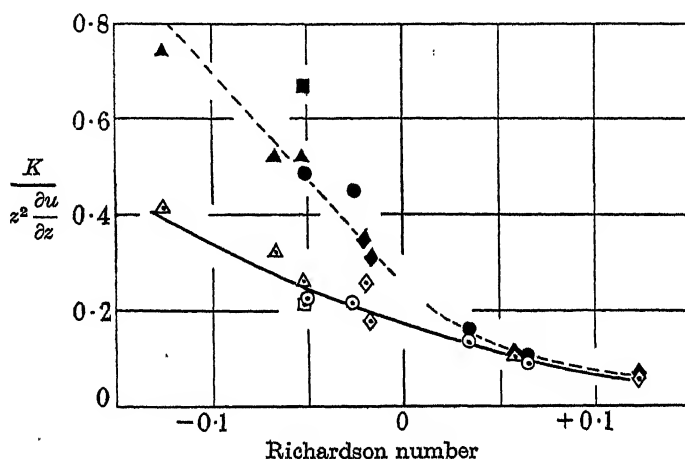


FIGURE 5. Observed values of eddy diffusivities for heat (K_H) and water vapour (K_v) in relation to wind velocity gradient ($\partial u/\partial z$) and Richardson number (Ri) at 75 cm. over a short grass surface.

group	$K=K_v$	$K=K_H$
A	○	●
C	△	▲
C'	□	■
D	◇	◆

($z=75$ cm.)

limited data the error in the present computations would amount at most to an overestimation of the order of $0.01 \text{ cal./cm.}^2\text{min.}$ in the upward flux of long-wave radiation and hence a corresponding *underestimation* in the upward turbulent flux. Again, Robinson (1947, p. 146) has observed that the computed downward flux of atmospheric radiation may be too high. This would lead to an overestimation of the turbulent heat flux. However, the indications of Robinson's measurements are that this error could be of the order of $0.02 \text{ cal./cm.}^2\text{min.}$ These features are clearly insufficient to modify the higher values of the turbulent heat flux (c. $0.20 \text{ cal./cm.}^2 \text{ min.}$) to the required extent.

The above considerations of the accuracy of estimation of the turbulent heat flux provide no grounds for suspecting that the observed differences in K_v and K_H are other than genuine and of the correct order of magnitude. Although it is not proposed in this paper to deal in any detail with theories of turbulent mixing it may now be

noted that, in qualitative implication at any rate, the present results are not entirely unexpected. In a recent treatment of convection near the ground Sutton (1948) has assembled evidence for the existence of a difference in heat and momentum transport. Furthermore, Priestley & Swinbank (1947) have recently put forward a modification of the classical theory of turbulent transport of heat, taking into account buoyancy effects. The latter treatment possesses implications as regards the transfer of physical quantities other than that of heat, and provides reasons for expecting different values of the eddy diffusivity to apply to the transport of different properties. As far as is known the present results provide the first direct experimental demonstration of the existence of this feature in the cases of matter and heat.

9. APPLICATION TO THE INDIRECT EVALUATION OF NATURAL EVAPORATION

In conclusion, it is of interest to examine one immediate practical application of the results presented here. Before doing so a brief review of the essential conclusions is probably desirable. Measurements have been made of the factors which define the eddy diffusivity for water vapour, and a comparison made with corresponding measurements made elsewhere of the factors defining eddy diffusivity for momentum for flow in the absence of thermal stratification. Identity of these two diffusive properties of the atmosphere is thus directly confirmed. No such direct comparison has been possible for a thermally stratified atmosphere, but the results indicate the extent of the modifications imposed on the eddy diffusivity, and in so far as a recent functionally established law relating wind shear and shearing stress may be accepted explicitly the identity of K_v and K_m is shown to hold in unstable thermal stratifications, though not in stable stratifications. However, a rapid decrease of stability effect with decrease of height above the boundary is demonstrated. Finally, measurements of the turbulent heat flux, from heat-balance considerations, lead to values of K_H which are reasonably equal to those of K_v in stable conditions but are substantially and systematically higher in unstable conditions. No quantitative explanation can at present be offered for the latter feature, though qualitatively it is consistent with recent modifications of the theory of turbulent transport. Due regard should be given to the fact that these results have been obtained in a shallow surface layer of the atmosphere on a single well-exposed level site, and with this qualification in mind we may now consider their bearing on the problem of evaluating natural rate of evaporation from other more easily measured quantities.

The difficulty of performing direct measurements of natural evaporation as a routine procedure, as distinct from the special test measurements employed here, has led to two distinct meteorological approaches to the problem. The 'heat-balance' method, initiated by Ångström (1920) with reference to evaporation loss from lakes, and applied recently to land evaporation by Penman (1948), depends fundamentally on the assumption of identity in K_H and K_v . Thus in our nomenclature the heat-balance equation may be written

$$R_s = R_r + R_z + S - K_v \left[\rho c_p \left(\frac{\partial T}{\partial z} + \Gamma \right) + L' \frac{\partial \chi}{\partial z} \right], \quad (18)$$

where L' is the latent heat of vaporization, so that with the appropriate measurements or estimates K_V and hence $-K_V \frac{\partial \chi}{\partial z}$, i.e. the rate of evaporation, may be derived. The advantage claimed for this method is that except for the above assumption no knowledge of the processes of turbulent transport is required. The present results only support this assumption in stable conditions. In unstable conditions, for which values of K_H approximately double those of K_V are indicated here, considerable error could arise in the computed rate of evaporation. It is easily seen from the above equation that the error depends on the relative magnitudes of the vertical gradients of temperature and absolute humidity, and that when the former is 10^6 times the latter, a factor which was frequently attained and occasionally exceeded in the present observations, the computed rate of evaporation would be 33 % too high. Thus on the present indications the method may only be applied with confidence when the vertical temperature gradient (measured upwards) has a positive or a relatively small negative value, and is likely to be seriously in error for large negative gradients of temperature, when, other factors being constant, evaporation occurs at a maximum rate.

The 'hydrodynamical' method has been used by Sverdrup (1936) and others for the estimation of evaporation from the sea and applied recently by Thornthwaite & Holzman (1942) and Penman (1948) to land evaporation. Thornthwaite & Holzman's method assumes identity of K_V with K_m computed from a wind-profile law of the form of equation (1). The validity of this assumption is established here for air flow over a land surface in the absence of thermal effects. For convenience of application we may note that integration of equation (5) over the height range z_1 to z_2 , with K_V of the form given for K_m in equation (2), leads, on rearrangement in terms of equation (1), to

$$E = \frac{k^2(\chi_1 - \chi_2)(u_2 - u_1)}{\left(\log_e \frac{z_2}{z_1}\right)^2}, \quad (19)$$

which is the equation applied generally by Thornthwaite & Holzman. Evaluation of the rate of evaporation thus requires the measurement of wind speed and absolute humidity at two heights, the value of the aerodynamic constant k being 0.4. In unstable conditions this equation no longer holds, since K_V and K_m are now given by equation (11). From this we have, by a similar process of integration,

$$E = \frac{(1 - \beta)^2 k^2 z_0^{2(1-\beta)} (\chi_1 - \chi_2)(u_2 - u_1)}{(z_2^{1-\beta} - z_1^{1-\beta})^2}, \quad (20)$$

which is again evaluated by measurements of wind speed and humidity at two heights, z_0 being obtained from equation (1) and measurements in *neutral conditions*, β from equation (12) as described in § 6. The indications of the results discussed in § 5 and at the end of § 6 are that the above equations may be used to prescribe the rate of evaporation over short periods (1 hr. or more) with an accuracy usually within ± 20 %, while over extended periods some improvement in this could reasonably be expected. In stable conditions neither of the above expressions can be supported

by experimental data, and at present it is necessary to appeal to the observed decrease of stability influence as the boundary is approached and to apply equation (19) as an approximation, with measurements made preferably in the first $\frac{1}{2}$ m. above the surface. Since in general the rate of evaporation from a land surface in stable conditions of air flow may be expected to be of a low order, the error incurred by this approximation in estimating daily or longer period evaporation loss will rarely be of importance.

It should be noted that if the height of the vegetation cover is not very small compared with the heights above the soil surface at which wind speed and humidity are measured, the wind-profile laws are only satisfied by applying a 'zero-displacement' correction (see, for example, Deacon 1949), which is attributed to the fact that the boundary from which turbulent transport is effective is at some distance above the soil surface according to the height and density of the vegetation. Thus equation (1) becomes

$$u_z = \frac{1}{k} \sqrt{\left(\frac{\tau}{\rho}\right)} \log_e \frac{(z-d)}{z_0}, \quad (21)$$

and equations (10), (19) and (20) are similarly modified in the z term. When, as is usual, it is not possible to estimate d directly with sufficient precision, it is necessary to measure the wind speed at *three heights* in neutral conditions and to evaluate d graphically from the equation

$$\frac{u_2 - u_1}{u_3 - u_1} = \frac{\log(z_2 - d) - \log(z_1 - d)}{\log(z_3 - d) - \log(z_1 - d)}, \quad (22)$$

which follows from equation (21).

Finally, it is desirable to emphasize that the demonstration provided here of the superiority of the present hydrodynamical approach over the classical heat-balance method is contrary to conclusions drawn by Penman (1948). It should be noted, however, that while favouring the latter method Penman finds that it overestimates the daily free-water evaporation loss in midsummer, a result which would be expected from the present analysis. Furthermore, the hydrodynamical approach which Penman concludes to be inadequate differs fundamentally from the present method and is founded essentially on an empirical modification of a treatment which is basically justified only in the case of a so-called 'aerodynamically smooth' surface, a condition which is recognized to be the exception rather than the rule over a natural land surface (see, for example, Sheppard 1947, p. 217, for a brief discussion of this feature).

Acknowledgement is gratefully made to the Head and members of the School of Agriculture and the Director of the University Farm, Cambridge, for facilities provided during the course of this work, to the Director of the Meteorological Office for permission to publish the results obtained and to Mr J. E. Skilling of the Meteorological Office for assistance in the observational and analytical work.

REFERENCES

- Ångström, Å. 1920 *Geogr. Ann., Stockh.*, **2**, 237.
- Brunt, D. 1939 *Physical and dynamical meteorology*, 2nd ed., Cambridge Univ. Press.
- Calder, K. L. 1949 *Quart. J. Mech. App. Math.* **12** (in press).
- Deacon, E. L. 1949 *Geophys. Mem.* (to be published).
- Ertel, H. 1933 *Met. Z.* **50**, 386.
- Holzman, B. 1943 *Ann. N.Y. Acad. Sci.* **44**, 13.
- Nikuradse, J. 1933 *Forschungsh. Ver. dtsh. Ing.* no. 361.
- Pasquill, F. 1943 *Proc. Roy. Soc. A*, **182**, 75.
- Pasquill, F. 1949 *Quart. J. R. Met. Soc.* **75**, no. 325 (in press).
- Penman, H. L. 1948 *Proc. Roy. Soc. A*, **193**, 120.
- Prandtl, L. 1932 *Beitr. Phys. frei Atmos.* **19**, 188.
- Priestley, C. H. B. & Swinbank, W. C. 1947 *Proc. Roy. Soc. A*, **189**, 543.
- Robinson, G. D. 1947 *Quart. J. R. Met. Soc.* **73**, nos. 315-316, pp. 127-150, and unpublished results.
- Rossby, C. G. & Montgomery, R. B. 1935 *Pap. Phys. Oceanogr.* **3**, no. 3.
- Schlichting, H. 1936 *Ingen. Arch.* **7**, 1.
- Sheppard, P. A. 1940 *J. Sci. Instrum.* **17**, 218.
- Sheppard, P. A. 1947 *Proc. Roy. Soc. A*, **188**, 208.
- Sutton, O. G. 1934 *Proc. Roy. Soc. A*, **146**, 701.
- Sutton, O. G. 1947 *Quart. J. R. Met. Soc.* **73**, nos. 317-318, pp. 257-281.
- Sutton, O. G. 1948 *Quart. J. R. Met. Soc.* **74**, no. 319, pp. 13-30.
- Sverdrup, H. U. 1936 *Ann. Hydrogr., Berl.*, **64**, 41.
- Taylor, G. I. 1915 *Phil. Trans. A*, **215**, 1.
- Taylor, G. I. 1917 *Proc. Roy. Soc. A*, **94**, 137.
- Taylor, G. I. 1931 *Rapp. Cons. Explor. Mer*, **76**, 35.
- Thorntwaite, C. W. & Holzman, B. 1942 *Tech. Bull. U.S. Dep. Agric.* no. 817.

The Royal Greenwich Observatory

BY SIR HAROLD SPENCER JONES, F.R.S., *Astronomer Royal*

(*Lecture delivered 27 January 1949. Received 27 January 1949*)

[Plates 1 to 4]

The Royal Observatory was established by King Charles II in the year 1675 for the specific practical purpose of 'rectifying the tables of the motions of the heavens, and the places of the fixed stars, so as to find out the so-much-desired longitude of places for the perfecting the art of navigation'. At that time the most accurate star catalogue available was the catalogue of 1000 stars, prepared by Tycho Brahe about 1598; only 777 of the stars had been properly observed, and the star places, whose average errors were of the order of $1'$ to $2'$, were not sufficiently accurate for the purpose of determining longitudes. The best tables for giving the position of the Moon were liable to errors as great as $20'$. Flamsteed, who was appointed the first Astronomer Royal, realized that a good stock of observations, continued for many years, was needed in order to provide star places with all the accuracy that was attainable and to furnish positions of the Sun, Moon and planets which could serve as the basis for the construction of satisfactory tables of their motions. So, from its very foundation, the Observatory started upon systematic and long-continued programmes of observation which, throughout its history, have formed its most significant and important contribution to astronomy.

MERIDIAN ASTRONOMY

Until after the middle of the eighteenth century, when William Herschel embarked upon his studies of the structure of the sidereal system and upon his observations of star clusters, nebulae and double stars, astronomical observation was concerned almost entirely with the positions and motions of the heavenly bodies. But long after Herschel's pioneer work had opened up new fields of investigation, the observations at Greenwich continued to follow closely the lines originally laid down. There were improvements, of course, in the design and construction of instruments and in their optical quality; there were progressive refinements in technique and progressive improvements in accuracy, which in turn opened up new fields of investigation. The observations which Bradley made at Greenwich between 1750 and 1762, amounting to about 60,000 in all, were of a higher accuracy than any made previously and are, in fact, the earliest observations which are precise enough to be of use to the astronomers of to-day. Bradley was particularly careful in examining the errors of his instruments and in keeping the instruments in the best adjustment; he was the first to introduce corrections to atmospheric refraction for the temperature of the air and for the height of the barometer.

Halley in 1718 called attention to the fact that three of the bright stars, Sirius, Procyon and Arcturus, had changed their positions since Greek times, and that Sirius had perceptibly changed its position since the time of Tycho Brahe; the fixed stars were not, in fact, fixed, and the study of their proper motions added a new interest to the determination of stellar positions. Bradley himself discovered the two phenomena of stellar aberration and of nutation. William Herschel, by analyzing the proper motions of fourteen stars which had been determined with accuracy by Maskelyne at Greenwich was able to show that the Sun itself had a motion relative to the stars. The improvements in the tables of the motions of the Sun, Moon and planets, which resulted from the progressive increase in accuracy of observation and from the fact that observations were made all round their orbits, enabled the great Continental mathematicians of the eighteenth century, Euler, Clairaut, D'Alembert, Lagrange and Laplace, to prove that within the limits of accuracy of observation the movements of the bodies in the solar system could be accounted for in detail on the sole hypothesis of the Newtonian theory of gravitation.

These facts are mentioned to emphasize that the determination of the positions and motions of the Sun, Moon, planets and stars is work of fundamental importance in astronomy. It demands continuity of observations over a long period, made with the greatest care and precision. It was the work for which the Observatory was founded; it is the work which must always be its first concern, for it is work that is never completed but that will always go on. In 1875 Airy, after reviewing the work of 40 years at Greenwich since his appointment as Astronomer Royal, in the course of which he had expanded the work into various new directions, remarked:

'Turning now from the past to the future, I see little in which I could suggest any change. If it should ever be necessary to make any reduction, I should propose to withdraw Meteorology, Photoheliography, and Spectroscopy; not as unimportant in themselves, or as ill-fitted to the discipline of the Observatory, but as the least connected with the fundamental idea of our Establishment.'

Until well into the nineteenth century, meridian telescopes for the determination of position formed essential equipment of most observatories. Many amateurs even made regular meridian observations and provided significant contributions to positional astronomy. But as new fields of investigation in astronomy have been developed, meridian observations have been discontinued at most observatories. They make heavy demands on observing resources and involve much computational work. For this branch of astronomy, continuity of observations with the same instruments over a long period is of the greatest value. Such observations are therefore not well fitted for University or private observatories. It is proper that they should be undertaken almost entirely, as in fact they are to-day, by the national observatories. In recent years the systematic pursuit of meridian observations has tended to be restricted more and more in the northern hemisphere to three observatories, Greenwich, Washington and Pulkowa. During the war the Pulkowa Observatory was completely destroyed, though it is now being rebuilt on a larger scale, and meridian observations will continue to form an important portion of its work. At the end of 1940 meridian observations at Greenwich had to be dis-

continued for the first time in the history of the Observatory; the complete discontinuance was not for long, but until after the end of the war it was possible to carry on the observations only on a small scale.

The purpose of meridian astronomy is to provide a fundamental system of reference, defined by the positions and proper motions of a network of stars distributed with reasonable uniformity over the whole sky, together with the numerical value of the constant of precession. The observations are made from a moving and rotating earth and it is convenient to use the equator as a plane of reference and the equinox as a zero point. Both the equator and the equinox depend upon the motion of the Earth and the motion of its axis; they can therefore be determined by observations of the Sun which, in effect, fix the orbit of the Earth, or by observations of the bright inner planets, which involve the elements both of the orbit of the planet and of the orbit of the Earth. Observations of the Sun are peculiarly liable to errors of a systematic nature, while observations of the inner planets, showing perceptible disks and phases, are liable to personal errors which are different from those that affect star observations. There is the further difference that observations of the Sun and planets are made by day, while the star observations are mostly made at night.

The position of the pole can be determined from observations of circumpolar stars, and the position of the nadir can be determined from observations with a mercury horizon. The two together will fix the equator point. There are normally systematic differences between the equator point fixed in this way and the equator point deduced from Sun and planet observations. The circumpolar observations, above and below pole, are made 12 hr. apart; they may be affected by systematic diurnal effects, by errors in the corrections applied for refraction, by changes in the instrumental adjustments, by instrumental flexure, which is different for the two observations, and by errors in the adopted figure of the instrumental pivots. There are consequently many possibilities of errors of a systematic or quasi-systematic nature. When observations of stars in the equatorial belt made at a northern and at a southern observatory are compared, systematic differences are generally found. The observations can be brought into agreement by adjusting the coefficients of refraction used in the reductions of the observations at the two observatories, but the corrections so obtained are usually found to be quite inadmissible.

The observations made with the conventional type of transit circle are, in fact, liable to many sources of error which are difficult to control adequately. If the same stars are observed with two different transit circles and the derived positions are compared, it is found that, in addition to random errors of observation, there are systematic errors in both right ascension and declinations which vary with right ascension and also with declination. Magnitude error in right ascension, which affected observations made by the old methods of eye and ear and of hand-tapping, have been practically eliminated by the use of the impersonal micrometer.

The sources of error with the conventional transit circle are numerous. The errors of instrumental adjustment—of level, of azimuth, and of collimation—are continually varying. The variations are difficult to control adequately. If the meridian opening in the transit pavilion is narrow, refraction anomalies are inevitable; with

a wide opening, the instrument is fully exposed to the wind and to temperature changes. The errors of azimuth are particularly difficult to control, unless the site is one which permits of fixed azimuth marks of high stability. Corrections must be applied for any departure of the figures of the pivots from perfect cylindricality; the determination of the figures of the pivots with accuracy and with freedom from spurious elliptical terms is not easy. Instrumental flexure can be determined only in the horizontal position of the telescope, so that a law of variation with zenith distance must be assumed. The telescope is turned in the course of observation into all sorts of positions, with possibilities of displacements of objective or of the mechanical parts of the micrometer, which may be sufficiently small to escape easy detection but which can introduce serious errors of a systematic nature.

The difficulty of eliminating systematic errors of instrumental origin can be illustrated by the reversible transit circle of the Cape Observatory, which was designed by Sir David Gill with much thought and care. The telescope can be reversed in its bearings, giving two positions for observation, denoted by E and W; the object glass and eye-end are arranged to be interchangeable, giving two arrangements denoted by I and II. Four separate combinations are therefore possible. After allowing for the difference in horizontal flexure in the two conditions I and II, the difference in the declinations measured in the two conditions has a regular run amounting to $0''.4$ from dec. $+40^\circ$ to the south pole. The difference between the two positions E and W in both conditions has a range of $0''.18$, but the difference is systematically larger in condition II than in condition I, the extreme differences being $0''.3$ and $0''.2$ respectively. The mean of the observations in the four separate combinations is adopted as likely to provide the closest approximation to the truth.

Because of these systematic errors of instrumental origin, the proper motions of stars must be based largely upon series of catalogues observed over many years with the same instrument, whereby the instrumental peculiarities are to a large extent eliminated. The observations made at Greenwich are of special importance for this purpose, as the Airy transit circle, which defines the prime meridian of longitude, has been used continuously since 1851. With this instrument more than 650,000 observations have been made. The last three programmes consisted of the observations of all the stars down to about the 8th magnitude in the zones of declination 0° to $+32^\circ$, $+32^\circ$ to $+64^\circ$, $+64^\circ$ to $+90^\circ$, supplemented by fainter stars in regions of high galactic latitude, together with numerous observations of a large number of fundamental stars. Thus the whole of the northern sky has been covered. The proper motions of the stars have also been investigated. The Airy instrument does not accord with modern ideas for the construction of a transit circle; it is not reversible in its bearings; it is housed in a pavilion which has buildings on each side with a narrow opening and liability to refraction anomalies; the terrain is asymmetrical to north and south, with the possibility that the atmospheric refractions north and south of the zenith may not be equal. The instrument has now reached the end of its useful life. A new transit circle, reversible and housed in a pavilion of semi-cylindrical shape, with a wide aperture and a fairly symmetrical terrain to north and south, was installed shortly before the war. Before bringing this instru-

ment into regular use, it has been subjected to a lengthy series of investigations, in the course of which the errors of the circle graduations and of the figures of the pivots have been determined with great accuracy. These investigations have thrown much light upon various ways in which the observations can be affected by systematic or quasi-systematic errors of instrumental origin.

These remarks have been made to illustrate the difficulties and complexities of meridian astronomy and to explain why observations of position must still be continued. The observations made at different times and with different instruments at different observatories are combined, by a process of adjustment which is to a large extent empirical, to form what is termed a fundamental system, so providing a system of reference that is more accurate than it is possible to obtain from observations with a single instrument and, moreover, covering the whole sky. The more effectively the various sources of error can be eliminated or controlled with each particular instrument, the more reliable the fundamental system will be. Both the places of the stars and their proper motions, as given in the fundamental system, are affected by errors; the errors of the star places increase with the lapse of time from the epoch of the system, so that successive revisions of the system become necessary. A process of gradual approximation, which still continues, is involved.

With the realization of the serious difficulties involved in the use of a large movable telescope for the accurate determination of positions, it is natural that consideration should be given to alternative designs of instruments in which the movable parts are reduced as much as possible. The use of a fixed telescope in conjunction with a moving plane mirror has more than once been suggested; a design of such a mirror transit circle has been developed at Greenwich. It is proposed to employ two fixed horizontal telescopes, with their axes in the meridian and their objectives facing a movable mirror, whose plane is parallel to the east-west axis of rotation. Stars would be observed in one or the other telescope according to whether they transit north or south of the meridian, while the two telescopes would serve also as collimators. Such an arrangement reduces moving parts to a minimum. The fixed telescopes can be of longer focal length than is convenient for a movable telescope, and they can be effectively insulated against rapid temperature changes. Observations of nadir, of level and of near-zenith stars can be taken in either telescope. Flexure effects are reduced to a minimum, while troubles arising from small displacements of micrometer parts or of objectives are entirely eliminated. The adaptation of a variable-speed motor drive to the micrometer wire is simplified. A small model of the proposed design has been constructed and the theory of the instrument has been investigated. The instrument appears to have great possibilities and it is hoped to try it out in practice.

Closely related to the meridian astronomy is the determination of the variation of latitude caused by the movement of the Earth's axis of rotation relative to the Earth. The pole has an irregular motion within a circle of about 30 ft. radius; it contains two principal components, of periods 12 and 14 months, but the motion is not sufficiently regular to predict ahead. The component of the motion along the meridian of a place causes a change of latitude; the component in the perpendicular direction affects time determinations. The complete motions can be determined by

observations of the variations of latitude at two places whose longitudes differ by about 90° . The variation of latitude was first established in 1888; Chandler afterwards found it clearly exhibited in observations back to 1750. Determinations were first made at Greenwich with the Airy reflex zenith tube, but the observations were not entirely satisfactory. A long series of observations was commenced in 1911 with the Cookson floating zenith telescope, the observations being made photographically. Pairs of stars at nearly equal distance north and south of the zenith are observed at meridian passage; each star makes a trail across the plate, the telescope being rotated through 180° between the observations of the two stars. The separations of the pairs of trails are measured, and the variations in latitude are deduced from the changes in these separations. Meridian observations of declination are corrected for the variation of latitude. A new photographic zenith tube, now under construction, will be used, when completed, for the determination of the variation of latitude as well as for time determination. The study of the polar motions raises many interesting problems and provides a means for the determination both of the constant of nutation and of the constant of aberration.

TIME DEPARTMENT

The provision of a time service is a normal function of a national observatory and is closely related to the work of the meridian department. The right ascension of a star is the sidereal time of meridian transit of the star. In positional astronomy a selected number of the brighter stars in the equatorial belt, suitably distributed round the sky, are selected as 'clock stars'. The positions and proper motions of the clock stars, as used at Greenwich, have been derived from the long series of meridian observations and are progressively refined as the observations continue. Observations extending from 6 to 12 hr. serve to control periodic errors in the right ascensions of the clock stars. Using these adopted right ascensions, the observations of the clock stars determine the errors of the standard sidereal clock night by night. The right ascension of any other star is derived by correcting the observed sidereal time of its transit for the clock error, interpolated for the time of transit.

Before 1927 the time determined at the Royal Observatory was based upon observations with the Airy transit circle. In 1926 a world programme of longitude determinations, in which a large number of observatories in all parts of the world participated, was undertaken under the auspices of the International Astronomical Union. For this special programme, a small reversible transit circle was used at Greenwich, the telescope being reversed near the middle of each transit, thereby eliminating the correction for collimation error. For such a programme it is necessary to adopt a common system of star places, in order to ensure that the derived longitudes are as free as possible from systematic errors in star positions determined with different instruments; in this particular programme the star places in Eichelberger's *Fundamental Catalogue* were used. It was found that the clock errors derived from these observations had a much smoother run than the clock errors derived concurrently from observations with the Airy transit circle. The latter are affected by obscure instrumental errors, to which reference has already

been made. It was found, moreover, that when the smoother clock errors provided by the small transit observations were used for the reduction of the Airy transit circle observations, the derived right ascensions of the stars became discordant; but that when the more irregular errors given by the transit circle observations were used, the derived right ascensions became accordant. The instrumental peculiarities of the transit circle are evidently involved.

Since 1927, therefore, the time determinations have been based entirely upon observations with small transit instruments, which are reversed in the middle of each transit. By resolution of the International Astronomical Union the revised Auwers fundamental system, known as the FK3, is used for the system of star places. The apparent places of the stars are taken from the annual volume *Apparent Places of Fundamental Stars*, published by the Nautical Almanac Office.

Prior to 1923 the standard clocks employed in the Royal Observatory were regulator clocks with Graham dead-beat escapements. A Cottingham clock, fitted with a Riefler escapement, installed after the first World War, was expected to give a higher standard of performance, but failed to come up to expectations. The development of the Shortt free-pendulum clock introduced a new standard of precision in time-keeping. In this type of clock a master pendulum, mounted in an airtight case, exhausted to a pressure of about 1 in. of mercury, and placed in a constant-temperature room, synchronizes a slave clock of commercial type. The master pendulum swings freely, except when given a small impulse once each half-minute, and is relieved of the work of moving a train of wheels to show the time on dials and of sending out signals. The first clock of this type, Shortt no. 3, was installed at the Observatory in November 1924 and soon showed its superiority over other types of pendulum clock. Other clocks of this type were therefore installed and used both as sidereal and as mean time standards in the Observatory.

The introduction of the new standards resulted in an important change in the system of time employed. The transit of the first point of Aries or the vernal equinox defines the beginning of the sidereal day, 0 hr. sidereal time. But the precessional motion of the true equinox is not uniform, being affected by irregularities, due to solar and lunar perturbations, which are known as *nutations*. In consequence the sidereal day varies slightly in length. If we imagine a point moving uniformly along the equator, with a motion equal to the mean motion of the true equinox, and so that its extreme distances from the true equinox on both sides are equal, we may term this point the *mean equinox*. The sidereal time determined by observation is *apparent sidereal time*. A *mean sidereal time* can be defined by reference to the mean equinox, in which all days are of equal length; it is obtained by subtracting the nutation from the apparent sidereal time. Apparent sidereal time was good enough before the introduction of the Shortt free-pendulum clocks; their superior precision made it necessary to introduce the concept of mean sidereal time, which has been universally adopted. The detailed study of the performance of the Shortt clocks at Greenwich, which proved their superiority over other types of pendulum clocks, stimulated their introduction into observatories in many parts of the world.

The Royal Observatory is responsible for the distribution of time to the public. The first steps in the distribution outside the Observatory became possible with the development of telegraphic communications. In 1852 an electric clock was installed at the Observatory, which transmitted a signal each day that caused a time-ball, on the offices of the Electric Telegraph Company in the Strand, to drop. In 1865 signals were sent hourly to the Electric and International Telegraph Company's office, whence they were distributed over the railway network of the country. After the telegraph system was taken over by the Post Office, in 1870, a complete system for the hourly distribution of time through the Post Office was gradually developed, which made Greenwich time widely available. A further step in the widespread dissemination of accurate time followed naturally upon the development of broadcasting. Two of the Dent regulator clocks were modified to run as synchronized clocks under the control of one of the mean time free pendulums, and were provided with a system of contacts to enable time signals to be transmitted automatically every quarter of an hour to the British Broadcasting Corporation. The signals were in the form of six dots at intervals of a second, the last coming exactly at the hour, the quarter, or the half hour. These signals are the familiar B.B.C. 'six pips' Greenwich time signal, which, since 5 February 1924, have been transmitted on all B.B.C. wave-lengths several times daily.

A further service, designed to be of value for navigation, was commenced on 19 December 1927. From that date radio time signals have been sent out on a frequency of 16 kc./sec. twice daily, at 10 and 18 hr. G.M.T., from the Observatory, via the Rugby wireless station. These signals, which last for 5 min., are of the so-called vernier type, spaced 61 to the minute, enabling the error of a chronometer to be accurately determined by observing the instants of coincidence between the signals and the ticks of the chronometer. Corrections to the times at which the signals were emitted are published by the Observatory at approximately monthly intervals, for use where higher precision is required, as, for instance, in survey operations. For the distribution of these signals a special 'diminished seconds' slave clock was installed, whose pendulum swings 61 times in a minute, and which is kept in synchronization by the mean-time master pendulum. The service has more recently been extended by simultaneous transmission of the time signals on several short wave-lengths. The introduction in 1936 by the Post Office of the 'speaking clock', designed by the Post Office engineers and constructed at the Post Office Research Station, which is automatically controlled by hourly time signals from the Observatory, has made accurate time continuously available.

It is necessary that the Observatory should keep abreast of developments in precision horology so that the time service can meet all demands for precision that are made upon it. After the development of the quartz crystal clock, it soon became evident that a new standard of precision in time-keeping had been reached. It was therefore decided to instal a clock of this type at the Observatory in order that some direct experience could be gained of its performance in comparison with the performance of the free-pendulum clocks. A clock, using a Dye-Essen ring crystal, was constructed under the supervision of the National Physical Laboratory and installed in 1939. The quartz vibrator was adjusted to have a frequency of 100 kc.

per sidereal second; demultiplier circuits provided an output with a frequency of 500 cycles per sidereal second, which was used to drive a phonic motor. Although the performance of this clock was not altogether satisfactory, experience showed that it was free from the small erratic changes of rate to which the pendulum clocks were liable and which made prediction uncertain; it was found also that it would be more convenient, when additional quartz clocks were installed, to employ a fundamental frequency of 100 kc. per mean solar second instead of per mean sidereal second.

The outbreak of war stopped for a time developments which had been planned for an improved time service. A skeleton time service was at once installed at the Abinger Magnetic Station as a safeguard against the possibility of the service from Greenwich being put out of action. Towards the end of 1940 when the frequent air raids made night observations impossible at Greenwich, the time service was moved in its entirety from Greenwich to Abinger. Shortly afterwards a second time service was installed at the Royal Observatory, Edinburgh, with the co-operation of the Astronomer Royal for Scotland. For the remainder of the war, the two time stations, at Abinger and Edinburgh, were in use; the two stations were connected by direct land line with teleprinter communication, enabling clocks at either station to be recorded at the other station.

Plans were proceeded with for the installation of quartz-controlled frequency standards. The first group of three clocks was installed in 1943. These clocks, which were constructed at the Post Office Research Station, were of the Post Office Group IV type, in which a GT-cut plate of quartz is maintained in oscillation at 100 kc. per mean time second, by a bridge drive circuit, which reduces to a minimum the effect of variation in the supply voltages. The crystals are mounted in thermostatically controlled ovens, and the temperature range permitted by the modified Turner circuit employed ensures that variations in frequency from this cause are small. Regenerative frequency dividers are used to provide an output at 1000 c./sec., which can operate phonic motors, provided with contacts from which signals can be taken off.

The installation has since been considerably extended and the time service is now based upon six groups of three quartz crystal clocks, four of the groups being at Abinger, where the main time-service station has remained, and two at Greenwich, enabling a time service of much higher precision to be provided than was possible with pendulum clocks. The quartz clocks have the further advantage that relative errors and rates can be obtained much more readily and with much greater accuracy than with pendulum clocks. For measuring time and frequency differences, decimal counter chronometers are used, embodying scale-of-ten counter circuits. The 100 kc./sec. output from one of the primary standards is fed into the counting unit. A seconds impulse from one clock can be applied to start the count and a seconds impulse from another clock to stop it. A single reading of the time difference between the two clocks is obtained, being shown on five dials reading successively in units of tenths, hundredths, thousandths, ten-thousandths, and hundred-thousandths of a second. To provide a check, the roles of the two clocks can be interchanged. Alternatively, by throwing over a switch, a series of successive readings may be obtained, when the results are added in the counter. This procedure is of value when

the intervals are nominally constant but are subject to small erratic variations, as when comparing a clock with a time signal; ten consecutive readings can then be obtained with advantage. For frequency comparisons, the nominal 100 kc./sec. outputs from two clocks are fed into a comparator, where the levels are adjusted to a standard value. The adjusted outputs are then combined and the resultant beats are shown on a meter. By means of a trigger circuit, a pulse is sent to the decimal counter when the beat-frequency voltage passes through zero. The intervals between the beats are thus accurately timed, enabling frequency differences to be determined with an accuracy of at least one part in 10^{10} . The quartz clocks are provided in addition with automatic beat counters; by automatic counting of the number of beats between each pair of oscillators in a 24 hr. period, the change in the time difference between each pair of clocks in the course of a day is recorded in units of 10^{-5} sec. The intercomparisons between each pair of clocks, in each group of three, provides an automatic check against incorrect action of any of the beat counters.

An additional advantage of the quartz crystal clocks is that though they are rated approximately to mean time they can serve all purposes. Separate sidereal time clocks and diminished seconds transmitters for the rhythmic time signals are no longer required. The phonic motors, driven by the 1000 c./sec. output from one of the clocks, can be adapted to these additional requirements. In the phonic motors used at the Royal Observatory, the rotor consists of a laminated iron ring with 100 teeth cut on its inside surface. The six-pole stator within the rotor has corresponding teeth cut on its pole faces. The motor runs synchronously at 10 revolutions per second, driving by gearing a commutator wheel at one revolution per second. A contact spring, bearing on this wheel, closes an electrical circuit for one-tenth of each second. In order to obtain sidereal seconds, a special gearing is used. The gearing ratio employed is $\frac{119}{114} \times \frac{317}{330}$, which is about four parts in 10^9 smaller than the correct ratio. This small error in the gear ratio is immaterial, for when the rate of the clock in mean-time milliseconds has been determined, the rate of the sidereal impulses, in sidereal-time milliseconds, can be at once inferred.

The rhythmic time signals consist of a long dash at the minute, followed by a series of dots spaced at intervals of $\frac{1}{61}$ min. These signals are derived from a contact drum, which is driven through a 60/61 gear from a phonic motor. In order that the signals may be sent out at the desired instants, a phasing adjustment is provided for the signal transmitter. There are day-to-day variations in the time lag introduced by the land-line joining the Observatory to the Rugby wireless station. Some test signals are transmitted a few minutes before the time signals themselves; these test signals are received and recorded at the Observatory and by comparing the times with those of the outgoing signals, the land-line lag is deduced. An adjustment to correct for the lag is made by rotating the phonic motor stator, thus advancing or retarding the phase of the rotor. A rotation of the stator housing by 360° advances or retards the contact time by one-tenth of a second.

The most recent phonic motor equipment installed at the Observatory provides a complete contact assembly for the control of all the time signals sent out, including

not only the rhythmic signals but also the B.B.C. 'six pips' time signals and the hourly signals for the control of the Post Office speaking clock.

The use of quartz clocks has resulted in a much improved precision in the time service. Their freedom from small erratic changes of rate makes accurate short-term prediction possible. This is of importance for the control of precision frequency standards, which can be checked against a 24 hr. time interval of high precision. The Rugby 10 hr. time signals are transmitted so as to give an accuracy in the 24 hr. interval between the signals on consecutive days which does not normally exceed 1 msec. The difficulties of accurate long-term prediction, which is needed to carry over periods during which no time determinations can be obtained, are much greater. The quartz crystal oscillators are subject to an ageing effect, which causes a drift in frequency, more rapid at first and gradually decreasing, though never, as far as present experience is a guide, completely disappearing. Time determinations extending over some months are needed in order to derive the frequency drift with the accuracy needed for prediction. But a complication is introduced by the motion of the Earth's poles, which causes small displacements of the meridian; the result is that a perfect clock, compared with absolutely accurate time determinations, would appear to have a small variable error, which at Greenwich can amount to about ± 25 msec. As the polar motion has two principal components, with periods of 12 and 14 months, an incorrect determination of frequency drift is inevitable unless the effects of polar motion can be allowed for. The motion of the pole along the meridian can be determined by observing the changes of latitude which result from it; the motion in the perpendicular direction can be determined only from observations of latitude variation at another observatory, differing by about 90° in longitude. The determinations of latitude variation at Washington, longitude 77° W, are communicated regularly to Greenwich and are used to correct for the effects of polar motion and thereby to derive more accurate values of the frequency drifts of the clocks. It is noticeable that the application of these corrections has appreciably smoothed the apparent errors of the clocks.

The introduction of quartz crystal clocks has demanded an improvement in the precision of the time determinations. The period of several months, which is required for a satisfactory determination of frequency drift with the relatively large errors inherent in the time determinations with the small transit instruments, could be much reduced if an appreciable reduction in the errors of observation were achieved. Much consideration has therefore been given in recent years to the design of new instruments for time determinations.

A photographic zenith tube, which is expected to reduce the probable error of a time determination to a few milliseconds, is now in an advanced stage of construction. This instrument is based on Airy's design of the reflex zenith tube at Greenwich, with modifications to adapt it for photographic observation due to F. E. Ross and incorporated in his photographic zenith tube, now in Washington. The essential principle is the employment of a zenith telescope, whose tube contains a mercury horizon to reflect the light and to bring it to a focus in the second Gaussian point of the objective, thereby making the observations practically independent of any error of level. The important modification introduced by Ross

was the inversion of the objective, placing the flint component uppermost, whereby, with an appropriate separation between the two components, the second Gaussian point is brought a few millimetres below the lower face of the crown component. The observations are made photographically, the photographic plate being mounted in the Gaussian plane. The instrument was designed originally for the measurement of the variation of latitude; the upper portion of the instrument, which carries the objective and the photographic plate, is in the form of a rotary, which can be turned through exactly 180° . If two exposures are made on a star, the rotary being turned through 180° between them, the separation between the two images in the direction of the meridian is twice the zenith distance of the star. In practice, exposures of finite length are given, the plate carriage being travelled along during each exposure with the speed of motion of the star image. If the two exposures are accurately timed and are approximately symmetrical about the instant of meridian transit, the time of transit can be inferred from the small relative displacement of the images in the direction perpendicular to the meridian.

The advantages of this type of instrument for time determination are considerable. The observations being photographic, personal equations are eliminated. Errors of level do not affect the observations; there is no collimation correction to trouble about; observations in the zenith are independent of azimuth error. As the instrument is fixed, the various sources of error to which a moving instrument is liable cannot occur. A longer focal length can be used than is possible with a moving instrument, with the advantage of a correspondingly greater scale. Observations are restricted to the zenith, where atmospheric transparency is highest and refraction effects are at a minimum.

The instrument which has been designed at Greenwich differs in a number of important respects from the Washington instrument:

- (i) It has a larger aperture (10 in.) and longer focal length (135 in.).
- (ii) A plain ball-bearing is used for constraint of the rotary, in place of conical bearing, in order to reduce friction and to facilitate construction.
- (iii) An autocollimation method is used as a criterion of the angle of reversal of the rotary.
- (iv) As a fixed axis of rotation is not required for (iii), a definite constraint in the horizontal plane is not needed. The two working orientations are each defined by a pair of stops instead of by a single stop.
- (v) Adjustments to the objective are provided for squaring-on and for coincidence of the nodal plane and photographic plate.
- (vi) Automatic reversal is accomplished by means of a system of wires which exert a pure torque on the rotary and therefore no tilting torque on the tube. The system is such that unidirectional rotation of the driving shaft is converted into reciprocating rotation of the rotary.
- (vii) The plate carriage is annular and the plate-holder mount is circular so that symmetry of diffraction pattern is secured. The carriage constraints are external to the aperture.
- (viii) Relative motion of the carriage and rotary is made to approximate to pure translation by means of a compensating system of flexed rods, which constrain the

carriage in the horizontal plane to which the motion is restricted by means of three balls that roll between horizontal planes.

(ix) Uniformity of rate in the relative translation of carriage and rotary is obtained by a specially designed system comprising a differential roller and metallic tapes.

(x) The time scale is produced photographically by means of a clock-controlled lamp giving flashes of very short duration. An independent chronograph is not required.

(xi) The height of the mercury surface is accurately adjustable and, as criterion of adjustment for constancy of scale value, an optical null method has been introduced for use in conjunction with a suspended silica rod.

Some consideration has also been given to the design of a new type of transit instrument, designated as the Horizontal Transit Instrument. The essential feature is that the telescope system remains fixed (though adjustable) with its axis horizontal and in an east-west direction. The light from a star of any declination, near the position of meridian transit, is directed along the optical axis by a subsidiary optical system of constant deviation, which can be rotated about an east-west axis and can be set to the appropriate declination. The effect on time determination of its positional errors (whether due to maladjustment of the axis of rotation or to pivotal errors) is reduced to the second order. Level and azimuth errors of the telescopic system have the same effect on the observed time of transit as they do with the ordinary transit instrument; but since the telescopic system is not deliberately subjected to gross mechanical disturbances and suffers from no pivotal errors, these level and azimuth errors should be far more stable than in the reversible instrument. The collimation error is dealt with by duplication of the telescopic system and reversal of the subsidiary system, so that the essential advantage of the reversible instrument is not sacrificed. Observation is made at the common focal plane of the duplex telescopic system, from the two sides successively. The fixity of the telescopic system avoids errors due to flexure, and permits of the use of a focal length considerably greater than can profitably be used in the ordinary reversible instrument. The level is determined with reference to two mercury surfaces, one at each end of the instrument, by means of an autocollimation method.

Instead of following the star image with a movable micrometer wire, a variable-deviation system is used by which the light in the telescopic portion of the instrument is kept always axial as the direction of the incident starlight rotates. In this way the tolerances of certain essential adjustments are greatly increased. Further, this variable-deviation system acts also as a micrometer and as the means by which signals are sent to the chronograph. An additional advantage of this axial method is that the fiducial line that bisects the star image is not required to move in order to follow the star's image or to be linked to the signalling system as at present. Thus no mechanical errors are introduced at this point. An optical method is contemplated for defining the position of the variable-deviation system in such a way that in its performance as a micrometer or signal emitter the system will be effectively free from the effects of mechanical errors. A thorough examination of the theoretical aspects of the design has been completed.

MAGNETIC AND METEOROLOGICAL DEPARTMENT

The first extension of the work of the Observatory beyond that laid down in the Royal Warrant for its foundation came with the setting up by Airy in 1840 of a magnetic and meteorological department. Certain meteorological data are of importance for the astronomical observations; atmospheric refraction depends upon the barometric height and the temperature; atmospheric transparency, an important factor in photometric observations, is correlated with horizontal visibility; the measured variation of latitude depends to some extent upon the direction of the wind; the amounts of sunshine, of rainfall, and of clear sky at night give some indication of the general observing conditions. The astronomer has to make his observations at the bottom of a dense atmosphere, and it is only to be expected that atmospheric conditions can influence the observations in many different ways. The effects are often unsuspected and obscure in origin and may not be discovered until results are analyzed; it was, for instance, quite unsuspected in advance that the measured latitude would depend upon the direction of the wind, though not upon its velocity. With a complete record of meteorological data, the basic data are available for any purposes of subsequent analysis. The Observatory makes continuous records of wind direction and pressure, of the total flow of air, of dry- and wet-bulb temperatures, of barometric height, of rainfall, of sunshine by day and of clear sky at night—the last being recorded by the trails of Polaris and of δ -Ursae Majoris obtained with a small fixed camera pointing to the pole. Daily eye observations are made of the barometer, dry-bulb and wet-bulb thermometers, radiation and earth thermometers, of the amount of cloud and of visibility. Some of the instruments used, such as the anemometers, are not of the most modern type, but the long series of observations made according to a uniform plan and with the same instruments is of special value for climatology. The Greenwich series of observations does, in fact, hold a unique place in British climatology. The data are of importance for various statistical purposes, such as questions of public health, the occurrence of epidemics, etc.; the meteorological results are therefore sent weekly to the Registrar-General. Observations are communicated daily to the Meteorological Office.

Magnetic observations were commenced in 1840 at the same time as the meteorological observations. Though not closely related, the magnetic and meteorological observations have always been in the charge of one department of the Observatory. This was primarily a matter of administrative convenience, to keep the non-astronomical work separate from the astronomical.

When the magnetic observations were started, they were made visually every 2 hr. throughout the day and night, but on one day each month they were made at 5 min. intervals throughout the 24 hr. This severe and trying labour was eliminated by the introduction in 1848 of continuous photographic registration, which has been maintained ever since, though with various changes and improvements in the recording instruments. A century of photographic registration has therefore been completed; the records are stored at the Observatory and are of great value in a variety of investigations.

In 1923 it became necessary to remove the magnetic observations from Greenwich because of the plans for the electrification of the local railway system. A site was selected near Abinger, on the slopes of Leith Hill, in Surrey, where a new magnetic observatory was built and observations were commenced in 1924. At that time the absolute observations of horizontal intensity were made with the Kew magnetometer, and those of dip with the dip inductor, which had superseded the dip circles in 1913. A few years later coil magnetometers were introduced as the standard instruments for the absolute measurement of horizontal and vertical intensity. The Schuster-Smith coil magnetometer for the measurement of horizontal intensity was installed in 1927, and the Dye coil magnetometer for the measurement of vertical intensity in 1928. Both these instruments were constructed at the National Physical Laboratory and are on loan to the Observatory from the Laboratory. The potentiometers used in conjunction with them are checked from time to time at the National Physical Laboratory.

The Schuster-Smith coil magnetometer has proved greatly superior to the Kew unifilar magnetometer in both speed and accuracy. The speed of observation is particularly valuable when conditions are at all disturbed. The base-line values of the horizontal intensity magnetograph deduced from the absolute observations have an uncertainty of not more than 1γ . The scatter of the base-line values of the vertical intensity magnetograph, deduced from the absolute observations with the Dye coil magnetometer is a little greater, but the uncertainty is only about 2γ or 3γ . This instrument is adopted as the standard for vertical intensity, the dip being deduced from the observed values of the vertical and horizontal intensities.

Absolute observations of declination are made several times every weekday, using for reference an azimuth mark whose azimuth is controlled by observations of Polaris; those of horizontal and vertical intensity are made daily, except Sundays. Frequent observations of horizontal intensity are made with the Kew magnetometer and of dip with the dip inductor; these observations serve as a general check on the observations with the coil magnetometers and are not otherwise used.

The Royal Observatory was the pioneer in using electrical coil instruments as standards; it is of interest to remark that a small systematic difference between the dip inferred from these observations, and the dip measured directly with the dip inductor was traced to an unsuspected defect in the inductor, arising from slight play in the bearings of the rotating coil.

The recording variometers, which record declination, horizontal intensity, and vertical intensity, are of the well-known la Cour type. They include both slow-run and quick-run variometers. Records are also obtained with declination and horizontal intensity magnetographs of low sensitivity, which are of value in following the field changes during great magnetic storms when the large rapid movements cannot always be followed with certainty on the normal records. The published data include the hourly means of each element throughout the year, together with the monthly mean hourly values and the means for the five international quiet days and the five international disturbed days each month; the daily mean and daily extreme values for each element, with the corresponding monthly means for all days, for the quiet days, and for the disturbed days; the mean diurnal in-

equalities for each month, for the year, and for winter, equinox, and summer, of declination, dip, horizontal intensity and for north, west, and vertical components, for all days, for international quiet days, and for international disturbed days separately; the harmonic components of the diurnal inequalities of north, west, and vertical components, for each month, for the year, and for the three seasons, for all days, quiet days, and disturbed days; together with mean monthly and annual values for all elements.

The daily magnetic character figures, and the 3-hourly range indices are assigned on the basis of the daily records and are communicated regularly to the international centre at De Bilt. As opportunity offers, the estimation of the 3-hourly range indices is being carried backwards through the long series of Greenwich records, providing data of fundamental importance in many geophysical investigations.

The Royal Observatory has for some 30 years assumed the responsibility for the preparation of the world magnetic charts which are published by the Hydrographic Department of the Admiralty. Charts of declination are prepared at 5-yearly intervals; of horizontal intensity and of dip at intervals of 20 years. During the war, in connexion with the degaussing of ships as a protection against magnetic mines, a world chart of vertical intensity was prepared. It has been decided that charts of horizontal intensity, of dip, of vertical intensity, and of total intensity will be prepared in future at intervals of 10 years, in accordance with a recommendation of the Association of Terrestrial Magnetism of the International Union of Geodesy and Geophysics. The preparation of these charts involves the collection and examination of magnetic observations and surveys made in all parts of the world; from these observations the secular change and the rate of change of secular change have to be inferred in order to reduce the observations to a common epoch and to extrapolate to the epoch for which the charts are prepared.

Since the untimely loss of the non-magnetic ship, the *Carnegie*, put an end to the long series of ocean magnetic observations undertaken by the Department of Terrestrial Magnetism of the Carnegie Institution, Washington, the magnetic data over some of the ocean regions, and particularly over the southern Indian Ocean, have become increasingly uncertain. Reports received from vessels of the mercantile marine were sufficiently concordant to justify some empirical corrections to the charts. The matter was discussed with the Hydrographer of the Navy and, as a result of representations made to the Board of Admiralty, the construction of a non-magnetic ship was decided upon. The ship, known as the R.R.S. *Research*, was in an advanced stage of construction at the time of the outbreak of war, when work had to be suspended. The possibility of completing the ship, except for the auxiliary engines, and of putting her into commission as a sailing ship is under consideration, though no decision has yet been reached.

The harmonic analysis of the world magnetic charts and the comparison between the observed and computed field at various points on the earth's surface can give some indication of areas where the charts are seriously in error. The charts for 1922 and 1942 were analyzed in this way, and in each case it was found that the computed position of the north magnetic pole was not in agreement with the adopted position, which was determined by Amundsen in 1904 and was in close agreement

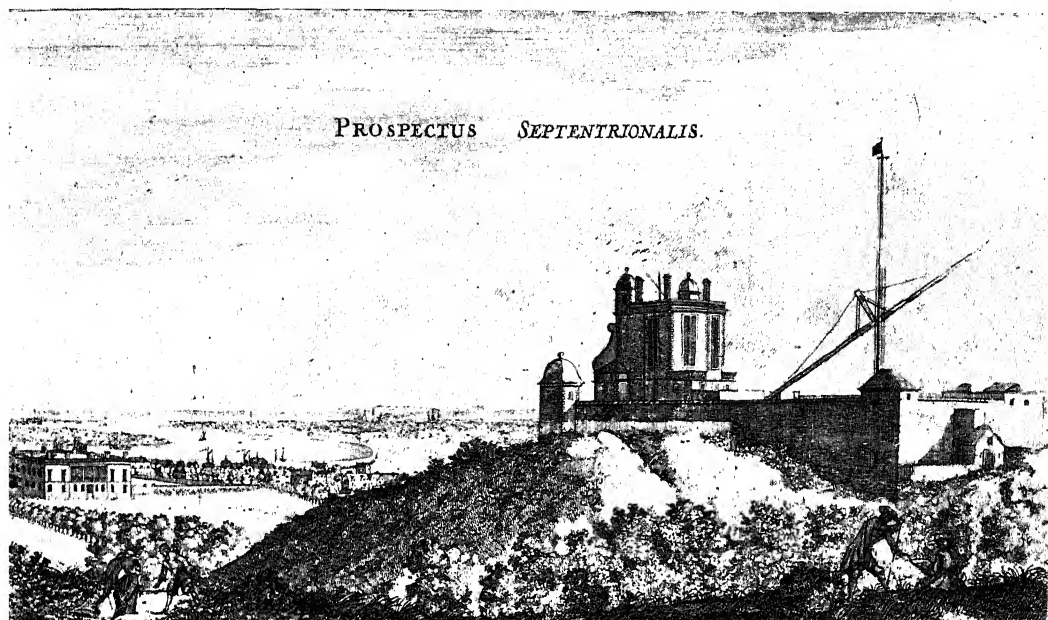


FIGURE 1. View of Flamsteed's original observatory (Wren building) looking north.

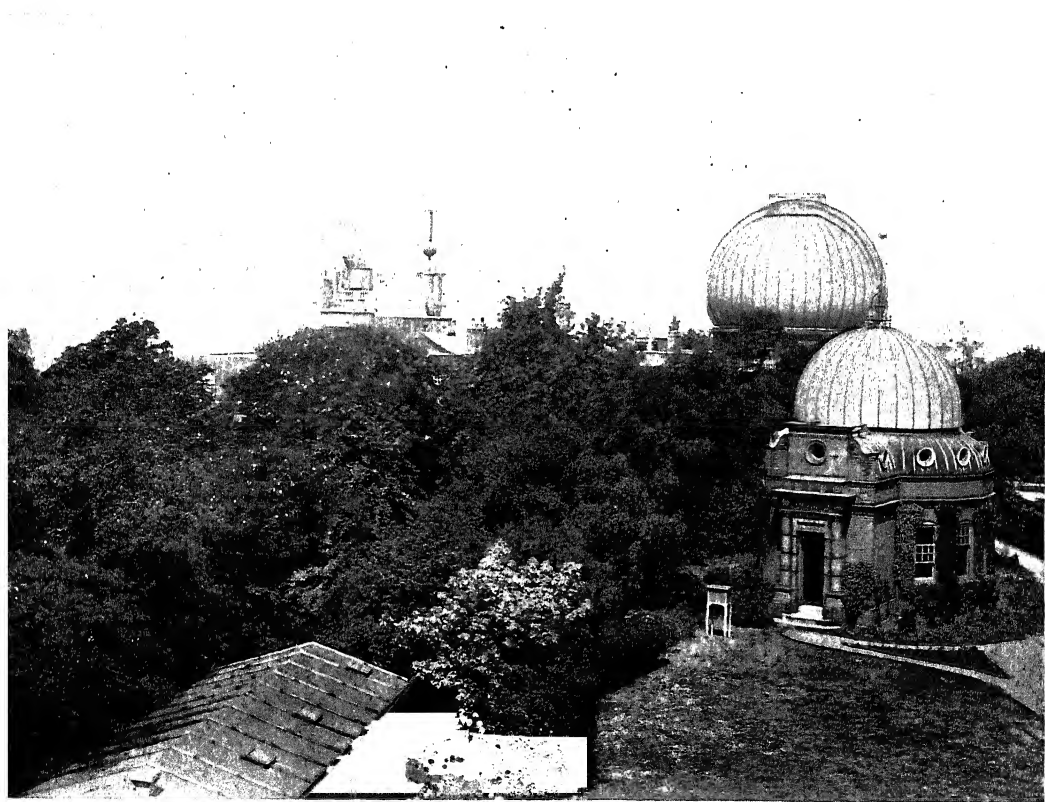


FIGURE 2. View from the 26-inch dome, looking towards the Wren building (circa 1930).

(Facing p. 156)

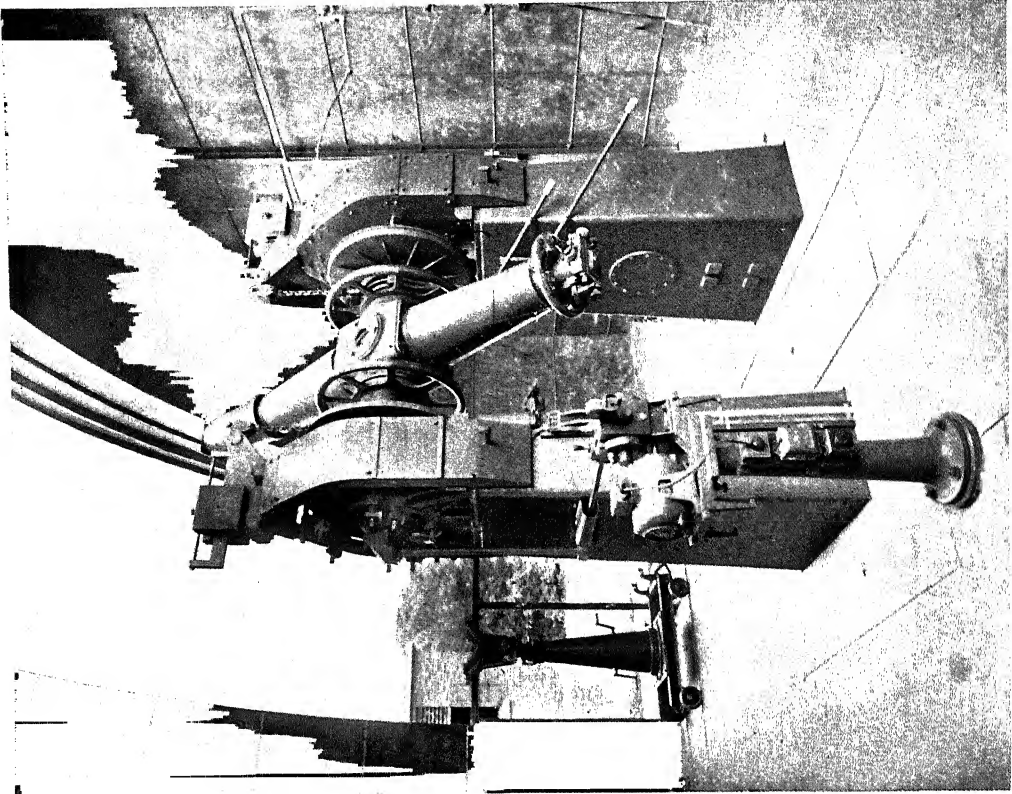


FIGURE 4. Reversible transit circle.

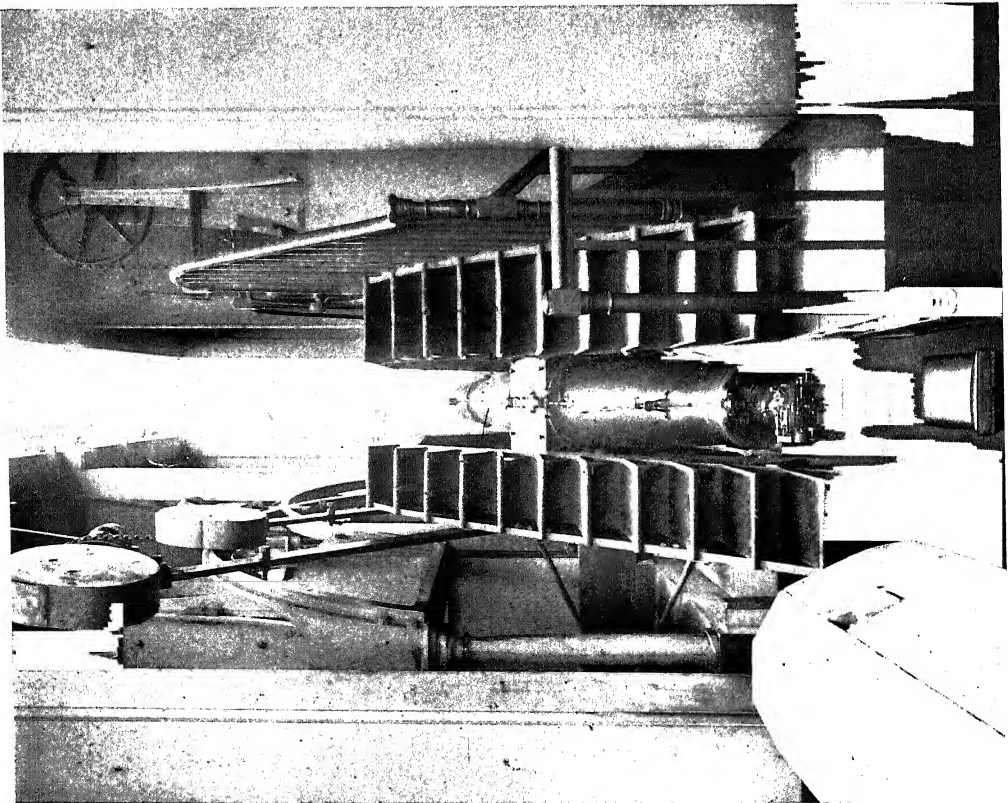


FIGURE 3. Airy transit circle.

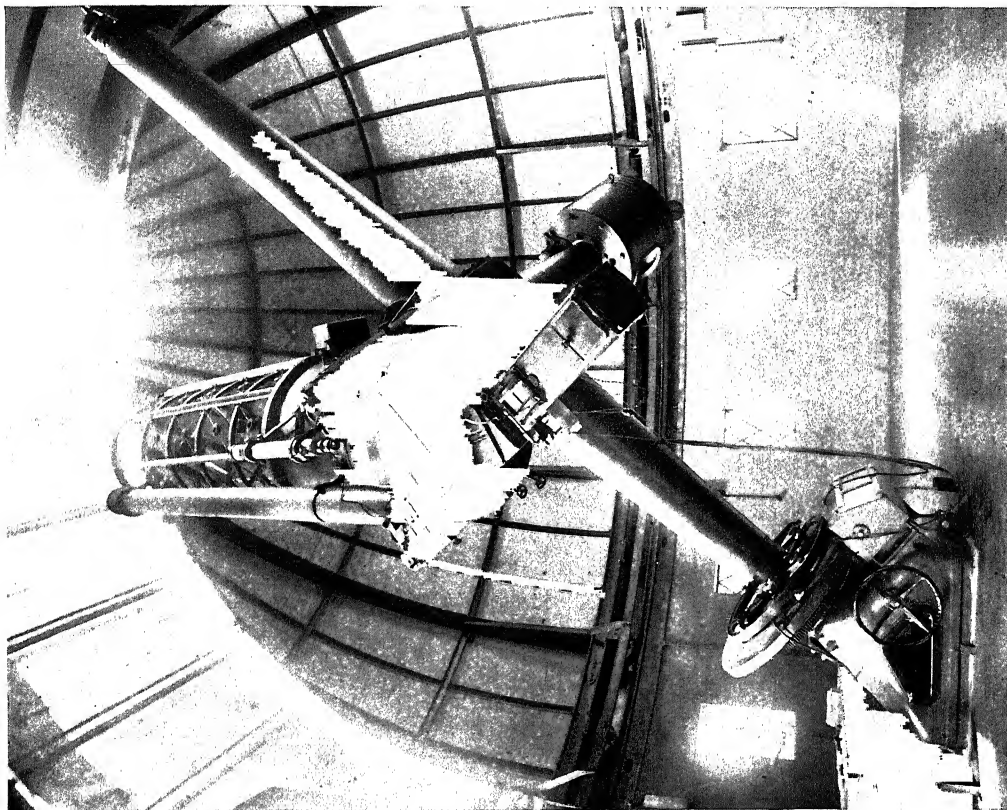


FIGURE 6. Yapp 36-inch reflector.

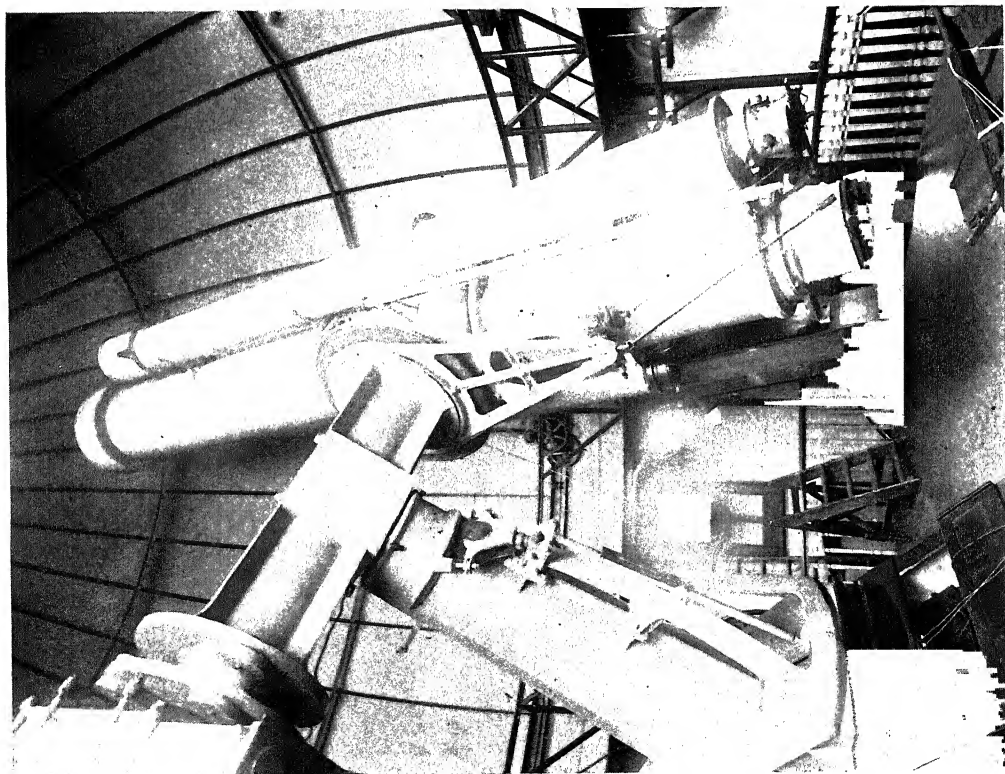


FIGURE 5. Thompson 26-inch refractor.

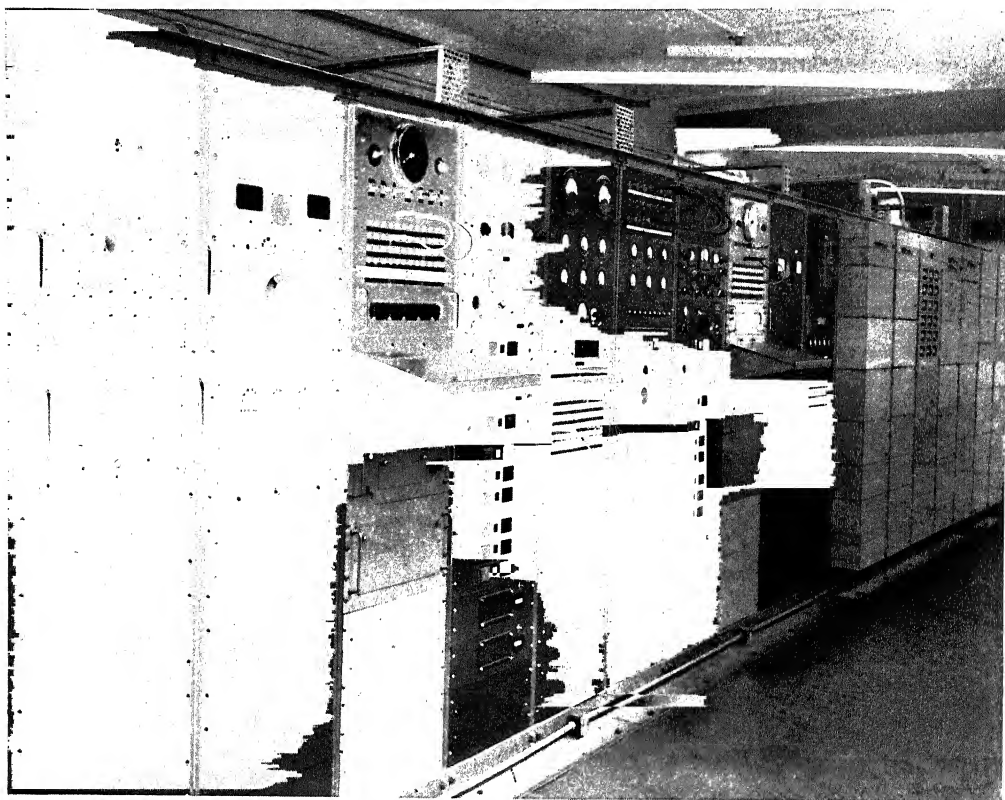


FIGURE 7. Control-room of Time Department.

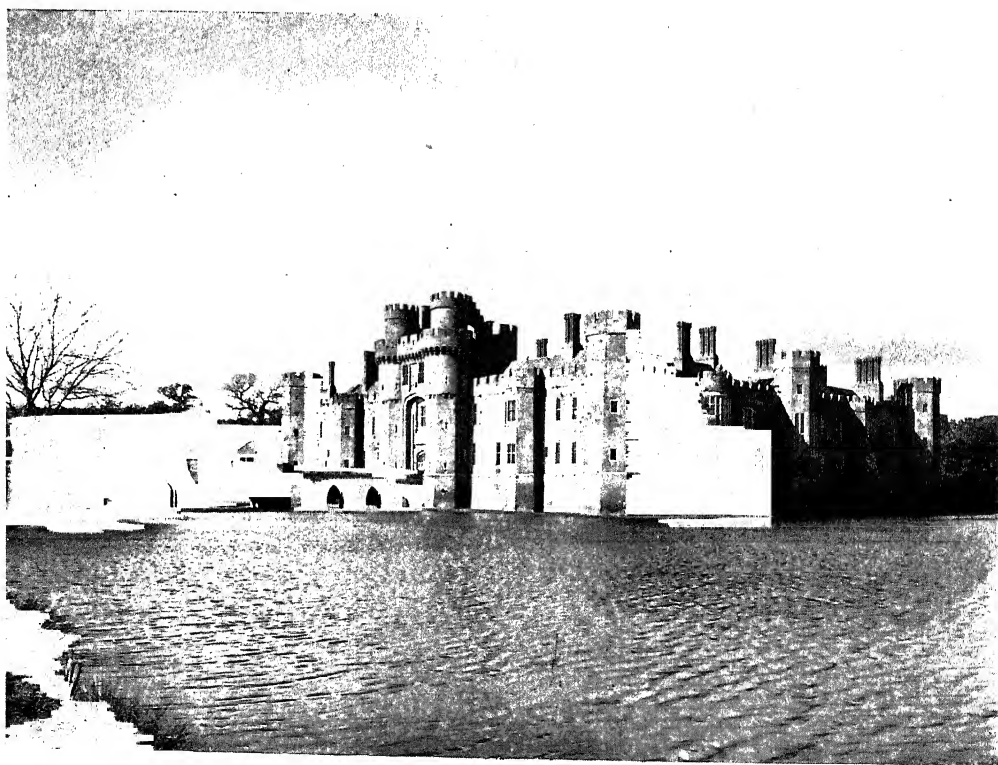


FIGURE 8. General view of Herstmonceux Castle.

with the position assigned by Ross in 1831. The discordance appeared to be too great to be attributable entirely to errors in the charts, and it appeared probable that the magnetic pole had moved appreciably from its position in 1904, in a direction somewhat to the west of north. When in 1945 a series of polar flights by the Lancaster aircraft *Aries* was being planned by the Empire Air Navigation School, there seemed to be an opportunity for obtaining some evidence on this question. The Commandant of the Empire Air Navigation School agreed that such observations as were possible in conjunction with the other objects of the flight should be undertaken. One of the flights actually made passed over the Amundsen position of the magnetic pole and another passed near the computed position. The results of these flights provided strong supporting evidence for the movement of the magnetic pole. More recent observations made by the Canadian Eastern Arctic Patrol have fully confirmed the displacement, though not by so large an amount as the harmonic analysis had suggested.

Since the termination of magnetic observations at the Kew Observatory, the Royal Observatory has taken over the responsibility for the testing and certification of magnetic instruments of different types. This work is undertaken not only for Government Departments and Colonial Governments, but also for various institutions and commercial firms.

THE SOLAR DEPARTMENT

The magnetic observations, which are of some importance for navigation, led to a further development, which has no connexion with navigation. About the middle of the nineteenth century it was discovered independently by Sabine, Lamont and Wolf that magnetic phenomena have a period similar to the 11-year sunspot period, which had been announced shortly before by Schwabe. It was also found that magnetic storms often occurred when there was a large spot near the centre of the Sun's disk. The relationship suggested the need for supplementing the magnetic observations by solar observations. The solar department was accordingly established by Airy in 1873. A photo-heliograph, in which the primary image of the Sun is magnified by an enlarging lens to give an image 8 in. in diameter, was installed, and the regular daily photography of the Sun, whenever conditions permitted, was commenced. These observations have been continued without interruption. Photographs with a similar instrument are made at the Cape Observatory and are forwarded to Greenwich. There are normally but a few days in the year which are not represented in the combined Greenwich and Cape series; photographs for the missing days can usually be obtained on request from either the Kodaikanal Observatory or the Mount Wilson Observatory. The positions and areas of sunspots and faculae appearing on each photograph are measured. A general catalogue of sunspots and ledgers both of recurrent and of non-recurrent groups are prepared from the measures. The total areas of umbrae, of whole spots, and of faculae for each day are computed, as well as the mean areas and heliographic latitudes for spots north of the equator, for spots south of the equator, and for all spots. The collected results provide the most complete information that is available about

sunspots and faculae, and one of importance in the study of the relationships between solar and terrestrial phenomena.

The sunspot data have been used at Greenwich in a variety of investigations. The position of the Sun's axis was determined from the observations of sunspots in the period 1874 to 1912, using both recurrent groups and groups observed for eight or more days; the data were analyzed for three complete spot cycles separately, for four different phases of the cycle, and for the three chief zones of heliographic latitude. The inclination of the Sun's axis of rotation to the ecliptic was found to be $7^{\circ} 10' \cdot 5$, and the longitude on the ecliptic of the ascending node to be $73^{\circ} 46' \cdot 8$ (epoch 1850.0). The motions of recurrent spots observed in five complete spot cycles, from 1878 to 1933, have been analyzed to determine the Sun's rotation period and its dependence upon heliographic latitude. The rotation periods derived from the five separate cycles were in excellent agreement, in contrast to the large secular change given by spectroscopic observations; the sunspots, however, because of their cyclic fluctuation in frequency, cannot be used to determine the rotation period year by year.

Many investigations have been made at Greenwich of the relationships between sunspots and terrestrial magnetic disturbances. The general statistical relationship between the occurrence of magnetic storms and the sunspot state of the central region of the Sun at the times of occurrence of the storms has been fully established. The largest storms tend to be associated with the largest spots; on the other hand, though the largest spots have a strongly marked tendency to persist for several rotations, the largest storms show little or no tendency to recur after 27 days—the period of the solar rotation—whilst moderate storms show a marked recurrence tendency. When a spot appears to be the source of a magnetic disturbance, the spot is usually situated, at the moment when the storm begins, between 2 days east and 4 days west of the Sun's central meridian.

In 1929 solar observations were extended to include visual observations of the Sun's disk in $H\alpha$ light, using a spectrohelioscope lent by the Mount Wilson Observatory. The initial purpose of these observations was to detect any special disturbances on the Sun that might be related to the occurrence of magnetic storms. Measurements are made, with the line-shifter, of the radial velocities of dark markings and, in particular, of those in the neighbourhood of sunspots. The intensities of bright $H\alpha$ flocculi and of prominences relative to the adjacent background are determined with a simple form of wedge photometer fitted with a comparison lamp. Visual measures of the contour of the normal Fraunhofer line $H\alpha$ at the centre of the disk are also made.

A special study has been made of the bright chromospheric eruptions or solar flares. The sunspot activity was on the wane when the spectrohelioscope was installed. By 1936 sunspot activity, having passed a minimum, was increasing rapidly. The number of flares showed a correspondingly rapid increase. By 1937, when the number of flares had still further increased, the association between the flares and sudden fadings of short-wave radio transmissions, more particularly in the case of the larger and brighter flares, had been fully established. The simultaneity of the two phenomena implied that the radio fadings were due to a solar agency

travelling with the speed of light. A direct comparison photometer to enable the peak intensities of the flares to be rapidly measured in relation to the adjacent continuous spectrum at 15\AA from $\text{H}\alpha$ was constructed in the workshop and installed in 1939.

The flares are found to occur mainly in the vicinity of large sunspots when in the stage of active development. In a number of instances the full sequence of phenomena has been observed; the brilliant eruption with the synchronous radio fade-out, accompanied by a typical bay or crochet on the magnetic trace, followed at an interval of the order of 1 day by a great magnetic storm. The solar influence on geomagnetic disturbance has led to the study of various geomagnetic phenomena. It is found that there is a marked diurnal variation in the times of sudden commencements, with a minimum at about 8 to 9 hr. G.M.T. A small proportion of sudden commencements have the initial movement in a direction opposite to the normal; these 'reversed' sudden commencements show an entirely different diurnal frequency, with a maximum at the time when the normal sudden commencements show minimum frequency.

The Brentwood radio station of Cable and Wireless Ltd. reports direct to the Royal Observatory any radio fade-out while it is in progress. Ionospheric data are sent regularly to Greenwich by the Superintendent, Radio Research Station, Slough, the Engineer-in-Chief, Radio Branch, G.P.O., and the Controller (Engineering), B.B.C. Information about sunspots and flares observed at Greenwich is supplied to various radio research centres, while an informal liaison over solar observations in general has been maintained with the Radio Group of the Cavendish Laboratory, Cambridge; the Radio Research Station, Slough; and the operational Research Group of the Ministry of Supply. It is of interest to note that though the solar observations were commenced at Greenwich purely because of the scientific interest in the possible relationship between solar phenomena and geomagnetic disturbance, the observations have become of practical value for the forecasting of ionospheric conditions. At the same time, they are of increasing importance for the theoretical investigation of the processes involved.

ASTROMETRY AND ASTROPHYSICS DEPARTMENT

In this department is included a wide range of investigations in astronomy which have developed from the application of photography to astronomy. It includes two main sections: (i) astrometry, involving precise measures of positions of star images on photographic plates, which may be purely differential, involving small displacements in position between two different epochs, or may be used to derive absolute positions, by using a number of stars as reference points whose positions have been separately determined by meridian observations; (ii) astrophysics, which is concerned with physical characteristics, brightness, colour, spectra, etc.

The first development of the work at Greenwich in this direction was the participation in the great international project, proposed in 1887, of a photographic chart and catalogue of the entire sky. A large number of observatories shared in this project, which included the preparation of a catalogue, giving the positions of

all stars down to a certain brightness, and the publication of star charts showing all the stars to a fainter limit of magnitude. To each observatory there was allotted a certain region of the sky, the Greenwich Observatory assuming responsibility for the cap of 25° radius around the north celestial pole. 13 in. photographic refractors of the same focal length, having a scale on the plates of 1 mm. to 1' of arc, each plate covering a field of $2 \times 2^\circ$, were to be used. The astrographic refractor for Greenwich was made by Grubb of Dublin, similar telescopes being made for a number of other observatories. For the determinations of position, all the stars in the Greenwich zones down to the limit of magnitude 9^m0 were observed with the transit circle in the years 1897 to 1905, while the measurement of the rectangular co-ordinates of the star images on the photographic plates was in progress. During these measurements, the diameters of the star images on the photographic plates were estimated in order to derive photographic magnitudes of the stars by the use of an empirical relationship connecting diameter and magnitude. In addition, the photographic magnitudes of all the stars brighter than 9^m0 were determined with a Cooke triplet lens camera of 6 in. aperture and 27 in. focus, covering a large field without appreciable distortion. Each region was photographed when at the altitude of the pole, the polar region being also photographed on the plate, so that the magnitudes of the field stars could be derived by comparison with the standard north polar sequence of magnitudes, which had been determined at the Harvard Observatory. The sequence of exposures was arranged so that the mean time of exposures on the field was in close agreement with the mean time of exposures on the pole, the assumption being made that the atmospheric absorption was equal under these circumstances for the two regions at the same altitude. This very large programme of work was spread over a number of years. The project was an ambitious one and proved to be beyond the power of some of the co-operating observatories, so much so that it has not even now been brought to completion after the lapse of more than half a century. The Greenwich section, both of the chart and of the catalogue, was one of the first to be completed. The results are published in rectangular co-ordinates, with sufficient data to enable the right ascension and declination of any star to be readily derived. For all the stars, however, down to magnitude 9.0 on the scale of the *Bonn Durchmusterung*, together with all fainter stars included in the catalogues of the *Astronomische Gesellschaft* and in Carrington's circumpolar catalogue, 16,780 stars in all, right ascensions, declinations, and photographic magnitudes were published in a separate volume.

Many programmes of observation undertaken at Greenwich since the completion of this work have been planned in order to make the available information about the stars in the polar cap, from dec. $+64^\circ$ to the north pole, more complete. The proper motions of the stars for which more than one position had been determined were derived by the comparisons of all available catalogues, each catalogue being reduced to a common basis by the application of systematic corrections. Then, commencing in 1923, the whole region was rephotographed with the astrographic telescope for the determination of relative proper motions of the stars. This second series of plates were exposed through the glass, so that by placing the corresponding plates of the two series of photographs film to film, the two images of each star were

brought into close proximity. It was necessary only to measure the small displacements between corresponding images differentially, and to apply corrections for differences in scale and orientation of the two plates in order to derive the relative proper motions. Comparisons between the photographic proper motions and proper motions based on meridian observations, where the latter were available, provided the data for converting from relative to absolute proper motions. The probable error of the derived proper motions was about $\pm 0''.8$ per century in each co-ordinate. The value of the *Greenwich Astrographic Catalogue* as a source of stellar positions was greatly enhanced by the determination of the proper motions.

The astrographic telescope has been used for a variety of other investigations. Mention may be made of a special determination of the magnitudes of the stars in the standard north polar sequence, using a coarse wire grating to give sensibly round first diffracted images, with a calculable difference in magnitude from the central image. The magnitudes of the sequence had been determined at the Harvard and Mount Wilson Observatories, but there was an appreciable difference in scale between the two determinations. The investigation at Greenwich proved that the Mount Wilson scale was correct. The telescope also co-operated, along with the 26 in. refractor, in the two international programmes of observation of Eros at the favourable oppositions of 1901 and 1931 for the determination of the solar parallax.

The main programme of observation with the 26 in. photographic refractor, which was presented to the Observatory by the distinguished surgeon, Sir Henry Thompson, has been the measurement of stellar parallaxes. Such observations demand great care and precision and, before the commencement of the work at Greenwich, had been made mostly in the United States with telescopes of much greater focal length. The high latitude of Greenwich is not favourable for stellar parallax work, because the short nights in the summer make it impossible to secure observations near the times when the parallax factors are at a maximum; more photographs are needed for the same weight in the parallax determinations than in lower latitudes. The weather at Greenwich, moreover, makes it difficult to obtain a proper balance in the observations at the several epochs at intervals of about six months which are required. By special care in the adjustment of the lenses of the objective, controlled by photographs at intervals to detect any tilt or eccentricity, and by other precautions, the results have proved to be of an accuracy comparable with that given by longer focus telescopes.

The observations have been confined to stars in the Greenwich astrographic zones. The observing lists include all stars in this region of magnitude 5.5 or brighter; stars down to 7^m with annual proper motions greater than $0''.10$; stars down to 8^m with proper motions greater than $0''.15$; fainter stars with proper motions greater than $0''.20$, together with stars of type K0 with proper motions greater than $0''.05$, the latter being included to obtain information about the distribution of these stars in absolute magnitude. The photographs are obtained with the use of one of a series of rotating sectors, the aperture being chosen to reduce the magnitude of the parallax star to about 11^m.5, the magnitude of the stars selected as comparison stars. For the brightest stars sufficient reduction in magnitude cannot be obtained by using a rotating sector; local desensitizing of the central region of

the plate with copper sulphate was tried for a time, in combination with a rotating sector, but the magnitude reduction produced by the desensitizing was uncertain. A neutral filter, giving a magnitude reduction of about 5^m was therefore employed, in combination with a suitable sector.

The observations were at first made by Kapteyn's method, in which the plate is exposed at one epoch, then stored undeveloped, and exposed again at the next epoch, the small displacements between the two series of images being measured. The method resulted in much loss of weight; good definition at one epoch might be followed by bad definition or exposures interrupted by cloud at the second. There were difficulties in balancing the different epochs. The method was therefore abandoned and each plate developed after exposure at the one epoch. For the measurement of the series of plates for the determination of the parallax of a star, suitable comparison stars were selected; a blank glass plate was then ruled with short fine parallel lines near the position of each comparison star, and of the parallax star. A plate was placed film down on the ruled plate, and the small displacements between each star and the rulings were measured. The ruled plate served as a dummy, which permitted of accurate setting of the micrometer wire and with which each stellar photograph was compared, thereby making possible the intercomparison between the stellar photographs themselves. Three separate exposures were normally given on each plate during the first series of observations. The practice was then introduced of giving two exposures, the plate being turned through 180° between the exposures; this procedure reduces any errors that may be caused by small local film distortions. It has been found that two exposures, with intermediate reversal, give the same accuracy as three exposures without reversal and with the advantage of saving time at the telescope. About 750 determinations of stellar parallax have been made since observations were commenced.

The 26 in. refractor has been used for a number of smaller programmes. Several series of photographs of Jupiter's satellites were obtained at the request of Professor de Sitter, to provide material for his determination of the elements of the orbits of the satellites and for his investigation into the theory of their motions. Photographic magnitudes of stars down to magnitude 14 in a number of Kapteyn's Selected Areas, zones +15°, +30°, +45°, and +60° declination, were determined by comparison with the north polar sequence; the magnitudes of 395 stars within 1° of the north pole were also measured.

A 30 in. reflector is mounted on the same mounting as the 26 in. refractor; the arrangement is not convenient, as it is never possible to use the two telescopes at the same time and there can be inconvenient competition between the demands on the same mounting for different programmes of work. The reflector has been used for the photography of comets and other celestial objects; in particular, a very fine series of photographs of Comet Morehouse 1908, whose tail changed markedly from night to night, was secured. On photographs taken with it, the eighth satellite of Jupiter was discovered by Melotte. The reflector was used also for a programme of determination of effective wave-lengths of stars in the north polar cap. Using a coarse wire diffraction grating over the end of the telescope, the distance between the two first diffracted images depends upon the grating interval, upon the focal

length and upon the wave-length of the light. The first diffracted images are really short spectra. The distance between the points of maximum photographic intensity determines the colour or effective wave-length of the star. The points of maximum intensity depend upon the distribution of light in the spectrum of the star, and upon the intensity curve of the photographic plate. The use of a reflector avoids difficulties from chromatic aberration. The growth of the diffracted images is not the same for stars of different colours, so there is an exposure-time effect, which must be allowed for. The image given by a 10th magnitude star with an exposure of 10 min. was used as standard. The change of effective wave-length with spectral type of the star depends upon the type of plate employed; with panchromatic plates and a yellow screen the change was found to be approximately linear, whereas with blue-sensitive plates there was little change between A0 and F8. The dispersion in the first diffracted images in this investigation was comparable in amount with atmospheric dispersion at a zenith distance of about 60° ; the results obtained throw much light on the relative displacements for stars of different types due to atmospheric dispersion, which may introduce systematic errors into the determination of the solar parallax from observations of the minor planet Eros or of any other asteroid.

In 1931, Mr William Johnston Yapp offered to present a large telescope to the Observatory. A 36 in. reflector was decided upon. The telescope was brought into use in 1934 for the continuation of a programme of observations of the colour temperatures of stars which had been commenced with the 30 in. reflector. The programme was a particularly difficult one to undertake at Greenwich, where the atmospheric transparency is poor, variable, and not uniform in amount in different directions. The colour temperature of a star is the temperature of a black body which has the same relative distribution of energy throughout the spectrum as the star. The determination of colour temperature involves comparison with a terrestrial source whose colour temperature is known; it is convenient, however, to divide the investigation into two parts, intercomparing the stars and then making comparisons with the terrestrial source. A selection was made of twenty-five stars, of spectral types A and B and fairly evenly distributed, to serve as standard stars. These standards were intercompared one with another, when at the same altitude. Other stars were then compared with one or more of the standard stars.

The reflector was employed with a slitless spectrograph. At first a coarse wire diffraction grating with dispersion at right angles to that of the spectrograph was used to provide a photometric scale. But this was wasteful of observing time. A scale spectrograph was therefore used with multiple slits, whose breadths were closely in the ratio of 1:2:4, the dispersion being approximately the same as that of the slitless spectrograph. The exact ratios of the light passing through the three slits were determined by a half aperture method. A Lummer-Brodhun cube microphotometer was used for measuring the spectral intensities, measurements being made at eight points in the blue and at eight in the red, free from absorption lines.

As standard source of comparison, a standard acetylene burner, with a nominal colour temperature of 2360° K, was used. The burner was specially calibrated at the National Physical Laboratory. The difference of colour from the stars was reduced by the insertion of a blue filter in the beam from the burner, whose absorption was

measured with the scale spectrograph. The burner was placed at a distance of about 600 ft. on the Octagon Room roof. The horizontal reddening of the acetylene flame in this distance was determined by special observations.

The colour temperatures of most northern stars brighter than $4^m.5$ and of spectral type A or earlier, as well as of many fainter stars of these types and of a selection of bright F- and G-type stars have been determined.

A slit spectrograph for use with the 36 in. reflector was completed in 1937. The optical parts are made of ultra-violet glass, and one-prism or three-prism dispersion can be used. The spectrograph was mounted towards the end of 1939, on the completion of the colour-temperature programme and various tests were made. But circumstances at that time made it impossible to commence any definite programme of observation. Since the war it has been employed in attempts to detect faint blue companions in late-type spectroscopic binary systems.

MISCELLANEOUS PROGRAMMES

Some items of the work of the Observatory do not come definitely within the scope of any particular department, but depend to some extent upon the personal interest of individual members of the staff. Expeditions have been sent from time to time to various parts of the world to make observations of total eclipses of the Sun, the programmes being determined by the problems of current importance. It was the expedition from Greenwich to Brazil for the observation of the total eclipse of 29 May 1919 which provided the first evidence to confirm Einstein's predicted displacement of stars in the vicinity of the Sun. The most recent expedition was a small expedition to Mombasa for the eclipse of 1 November 1948, to test a method of accurate determination of the position of the Moon, designed to be used on the occasions of a total eclipse at two widely separated centres, for the purpose of providing an accurate geodetic connexion.

The 28 in. refractor, installed in 1886, has been employed for many years for the observation of close double stars with a filar micrometer. The measures obtained up to 1919 were collected and published in a single volume. The observations were used, in conjunction with observations made elsewhere, for the determination of the dynamical parallaxes of 576 double stars. The parallax of a binary system of known period and angular dimensions of orbit can be calculated if the combined mass is known; in the absence of a knowledge of the mass, the error introduced by assuming the combined mass to be twice that of the Sun is relatively small, as the mass enters only to the power of $\frac{1}{3}$. The parallax so deduced is called the dynamical parallax. When the binary star has not been observed for a complete period, the parallax can be estimated, though with less certainty, from the rate of change of angle and distance.

A marked improvement in the accuracy of the measures followed from the introduction of a comparison image micrometer, constructed in the Observatory workshop. A Wollaston prism is used to give a double image of an artificial star, the separation of the two images being varied by altering the distance of the source from the prism. Rotation of the prism rotates the position angle of the artificial

binary. Crossed nicols permit the brightness of either image to be varied. A blue filter is used to give a colour temperature of about 5500° K, so that the artificial images appear of about the same colour as the stellar images. The position angle, separation and magnitudes of the two images are adjusted to be comparable to those of the binary star; the four images, two real and two artificial, are set to form the corners of a parallelogram. The eye is sensitive to any slight lack of symmetry, even under unsteady conditions when the use of a filar micrometer is difficult. The comparison image micrometer is particularly advantageous in the measurement of close pairs and, because no field or wire illumination is required, it enables fainter stars to be observed.

CHRONOMETER DEPARTMENT

In 1821, soon after the control of the Royal Observatory passed from the Master-General of the Ordnance to the Board of Admiralty, the charge of chronometers used in H.M. Navy was transferred to the Royal Observatory. In the following year public trials of chronometers were instituted. Makers were invited to submit chronometers for trial, with a view to purchase by the Admiralty, and money prizes of considerable value were offered for the best chronometers. This system of annual trials was continued up to the beginning of the World War of 1914–18, though the giving of special prizes was discontinued after a few years. These trials had a considerable influence in stimulating improvements in the design and construction of chronometers.

Because increasing difficulty was found in getting chronometers and navigational watches satisfactorily repaired and adjusted by the trade, a Repair Shop was started in connexion with the Chronometer Department in 1937, where repairers and adjusters could be trained to the high degree of skill essential for work on precision time-pieces. The Repair Shop proved invaluable during the war, when the numbers of repairs to be dealt with was very large and when the Chronometer Department was testing and issuing an average of from 25,000 to 30,000 chronometers and watches a year. It is expected that the amount of repair work undertaken will gradually expand, and a scheme has recently been introduced for the indenturing and training of apprentices. The facilities of the Repair Shop are available also for any fine precision work required in connexion with the construction and modification of instrumental equipment.

OPTICAL AND ELECTRONIC LABORATORIES

The progressive expansion in the scope of the work of the Observatory has involved the employment of much specialized equipment, which has had to be designed in the Observatory and made in the Workshop. The slitless spectrograph, the scale spectrograph, and the Lummer-Brodhun cube microphotometer, employed in the colour-temperature programme, and the comparison image micrometer for double-star observations are a few of the many items of equipment which have been made at the Observatory. The satisfactory design of new equipment not infrequently requires a considerable amount of preliminary investigations and tests, and the

need for a definite laboratory section has progressively increased. The introduction of quartz crystal vibrators as standard clocks required the use of much ancillary electronic equipment; because much of the special equipment needed was not available commercially, it became necessary to design and construct it in the Observatory. A laboratory section has accordingly been formed, comprising both an optical and an electronics laboratory. Started initially for the special needs of the Time Department at the time when the work of that department was expanding rapidly, the laboratories can meet the requirements of all departments of the Observatory. Amongst the electronic equipment which has been designed and constructed, mention may be made of a special radio receiver for the reception of radio time signals transmitted on the long-wave band with low carrier frequency, designed to secure constant lag, a high degree of selectivity, and a steep build-up curve, and with provision for selecting any particular point on the build-up curve and for securing constant output voltage; of a 2 Mc./sec. standard frequency transmitter, controlled by one of the 100 kc./sec. oscillators; of a receiver for comparing the frequency of the Droitwich 200 kc./sec. transmissions with any of the Abinger standards; of a number of double current electronic send relays; and of a great deal of switching, modulating, and monitoring equipment.

NAUTICAL ALMANAC OFFICE

In 1766, the year after his appointment as Astronomer Royal, Maskelyne published the first number of the *Nautical Almanac* (for the year 1767). The *Almanac* was designed for the use of seamen and particularly to facilitate the employment of the method of lunar distances for determining longitude. It proved to be a most valuable aid to navigation. Maskelyne continued to produce it annually for 44 years, until his death in 1811. The computations for the *Almanac* were performed in duplicate by computers, mostly working at home, and it acquired a very well-deserved reputation for high accuracy. In 1781 Maskelyne published a volume of *Tables Requisite to be Used with the Nautical Ephemeris*, which was in effect a handbook for use with the *Almanac*. Pond, who succeeded Maskelyne as Astronomer Royal, did not take the same interest in the *Nautical Almanac*, though he remained nominally responsible for it. The *Almanac* lost its reputation for accuracy, and eventually in 1831 a separate Nautical Almanac Office was established with its own Superintendent and having no formal connexion with the Observatory.

In 1937, the Nautical Almanac Office was again placed under the direction of the Astronomer Royal and became a branch of the Royal Observatory, though retaining its own identity and its Superintendent. The *Almanac* at that time was responsible for the production and publication of the standard edition of the *Nautical Almanac* and of the *Abridged Nautical Almanac*, designed for navigational purposes. Shortly afterwards the office undertook the production of an *Air Almanac*, adapted to the special requirements of air navigation. The airman does not need to know his position as accurately as the sailor, but, because of the high speed of modern aircraft, he requires to deduce his position with the minimum of delay after making the observations. To meet these requirements, the data in the *Air Almanac* are

presented in a special way, Greenwich hour angle being used instead of right ascension, and to a lower degree of accuracy than in the *Abridged Nautical Almanac*. Special *Air Navigation Tables* were also prepared in the office and published, for use with the *Air Almanac*, to facilitate the rapid derivation of the position of the aircraft.

In 1940 the publication (for the year 1941) of an annual volume of *Apparent Places of Fundamental Stars* was commenced. This volume gives the apparent places, at 10-day intervals, for most stars, but at daily intervals for close circumpolar stars, of the 1535 stars in the FK3 Fundamental Catalogue; the time determinations at all observatories are based upon these positions. The computations of the apparent places are shared by the United States, France, Germany, and Spain; the Nautical Almanac Office is responsible for the co-ordination of the work, for the collation of the data, and for the preparation and publication of the volume.

At the Conference of Commonwealth Surveys in London in August 1947, a strong desire was expressed for a special almanac to be prepared and published to meet the needs of topographical surveyors. Detailed proposals were therefore prepared in the office for a *Star Almanac for Land Surveyors* and have been approved. The first issue will be made in 1950 for the year 1951.

The experience gained with the *Air Almanac* has confirmed the advantages of the method of tabulation of data according to Greenwich hour angle, and has given rise to a desire for the revision of the *Abridged Nautical Almanac*. Various alternative arrangements of presentation of the data have been considered, and, after much detailed consultation with all classes of users, the final form has been settled. In its revised form, which will be issued in 1951 for the year 1952, the *Almanac* will tabulate Greenwich hour angle in arc directly, instead of right ascension.

Special investigations into methods and tables for both sea and air navigation have been made, including a comprehensive survey of tables for astronomical polar navigation. The office has also been responsible for the preparation and publication of various tables, including *Seven-Figure Trigonometrical Tables for Every Second of Time* (1939), *Five-Figure Tables of Natural Trigonometrical Functions* (1941), *Planetary Co-ordinates for the Years 1800–1940, Referred to the Equinox of 1950.0* (1933), and of a continuation volume for the years 1940–60 (1939).

The office has been a pioneer in the adaptation of computations, formerly performed by logarithms, to machines. Because of its wide experience in methods of numerical computation and its machine equipment, it was able to provide a computing service to deal with a great variety of problems for various Government departments, which presented themselves during the war. Much preliminary investigation was often needed to discover the best method of solving special problems, with the least expenditure of labour and of time. Approval has recently been given for the installation of a complete range of Hollerith-punched-card equipment, suitable for general computational work. It is intended to extend the use of the equipment, where suitable, to the work of the Observatory as a whole, as well as to the more routine work of the office. It is also hoped, eventually, to produce copy for some of the office publications automatically on card-operated machines.

The office has a close link with navigational problems and maintains a complete library of the navigational almanacs and tables of all countries. It is at present engaged on the computational work necessary for the latticing of charts required for the Decca system of navigation.

REMOVAL OF THE OBSERVATORY FROM GREENWICH

The conditions at Greenwich for astronomical observations have progressively deteriorated as London has grown outwards beyond the Observatory. The increasing pollution of the atmosphere and the increasing brightness of the sky at night have combined to affect adversely the quality and nature of the observations. Photometric and spectrophotometric observations, which require a uniform transparency in different directions and freedom from rapid variations of transparency, are practically impossible when clouds of smoke from nearby power stations and factories drift over the Observatory. But every type of observation is affected—meridian, solar, visual and photographic; in the exacting work of double star observations it has become impossible to observe close doubles which were observed with relative ease half a century ago. Under such conditions, the removal of the Observatory from Greenwich was essential if the Observatory was to continue to make useful contributions to astronomy. With the strong support of the Board of Visitors, proposals to remove the Observatory to a new site were submitted to the Board of Admiralty shortly before the outbreak of war. The war started before a decision was reached, and the question of removal had then to be deferred. During the war the principal instruments were partially dismantled, the Time Department was transferred to Abinger, and the Chronometer Department, with the Repair Shop, was moved first to Bristol and then to Bradford-on-Avon. The work of the Magnetic Observatory was becoming hampered by disturbances from the extension of railway electrification, and proposals for its removal to a site remote from railways were made. Widespread search for suitable new sites was carried out. After the termination of the war, the question of removal was again taken up. A short list of the most promising sites was prepared, and these sites were visited by a Committee of the Board of Visitors. Finally, it was announced on 11 April 1946 that Herstmonceux Castle in Sussex had been chosen as the future home of the Royal Observatory, and approval was given for the transfer of the Magnetic Observatory to a site to be selected in north Devon at a distance of at least 10 miles from any railway.

Some 372 acres of ground were acquired with the fifteenth-century castle, providing adequate space for erecting the various telescopes and for future additions to the equipment, and a safeguard against near encroachment by undesirable developments. A first stage of the removal is in progress. The Chronometer Department and the Secretariat have moved to Herstmonceux. A solar building to house the photoheliograph, two spectrohelioscopes, and spectrographic equipment for solar research is nearing completion. The transfer of the Solar Department, of the Magnetic and Meteorological Department, and of the Nautical Almanac Office should be possible during the course of the present year. Further stages of the removal, involving the erection of buildings and domes for telescopes, are under

consideration. The 26 and 28 in. refractors, whose domes were seriously damaged during the war, have been dismantled; the opportunity is being taken for some alterations to be made to these telescopes before re-erection takes place. The work of some departments of the Observatory is necessarily on a reduced scale during the present period of transition. There are, however, great hopes for the future, when the Observatory has been fully established in its new home and can reap the advantage of the good observing conditions. A selection of several possible sites in north Devon for the Magnetic Observatory has been made, and tests of their freedom from magnetic anomalies have been made. Some further sites will be examined before a definite selection is made.

On the occasion of the commemoration of the tercentenary of the birth of Sir Isaac Newton, held in London in July 1946, the President of the Royal Society announced that the Chancellor of the Exchequer had agreed to provide funds for the construction of a reflecting telescope of 100 in. aperture, to be associated with the name of Sir Isaac Newton and to be available for use by qualified astronomers from all observatories in Great Britain. It has been decided that the telescope will be erected in the grounds of the Royal Observatory at Herstmonceux. The telescope will be under the administrative control of the Astronomer Royal; a special Board of Management will be responsible for the scientific direction, including the designing of the telescope, the supervision of its construction, the consideration of programmes of observation, and the allocation of observing time between the various users of the telescope. The Board of Management will consist of the Astronomer Royal (Chairman), the Astronomer Royal for Scotland, and the Directors of the Cambridge and Oxford University Observatories as *ex-officio* members, together with four Fellows of the Royal Society and four Fellows of the Royal Astronomical Society. The telescope will enable British astronomers to undertake many programmes of observations which have hitherto been impossible because of the restricted light-gathering power of the largest telescopes at present in use in Great Britain, while the library, workshop, and other facilities of the Royal Observatory will be available to all users.

A note on the Riesz method and the method of residues

By F. C. AULUCK AND D. S. KOTHARI, *University of Delhi*

(Communicated by M. N. Saha, F.R.S.—Received 26 July 1948—

Revised 1 January 1949)

It is shown that the Riesz method of analytic continuation and the method of residues give the same results in the classical electromagnetic theory of a point source.

1. It is well known that in the framework of the classical electromagnetic theory, divergence-free results can be obtained by the λ -limiting process due to Wentzel, Dirac and Pauli, or by using the powerful method due to Riesz where the potential is obtained by analytic continuation to $\alpha = 0$ of any arbitrary parameter α . There is a third method in which the potential at any point is given by a contour integral (Frenkel 1926). In a recent paper Ma has established the equivalence, in the case of the field of a point source, of the λ -limiting process and the Riesz method. The present note deals with the equivalence of the method of residues and Riesz method. For the sake of completeness a brief description of both these methods is also given.

2. We define the matrix tensor $g_{\mu\nu}$ as

$$g_{00} = 1, \quad g_{11} = g_{22} = g_{33} = -1, \quad g_{\mu\nu} = 0 \quad (\mu \neq \nu).$$

The velocity of light is taken as unity. We write $[A, B]$ for the scalar product of two vectors A_μ, B_μ ,

$$[A, B] \equiv A_\mu B^\mu = A_0 B_0 - A_1 B_1 - A_2 B_2 - A_3 B_3.$$

The Maxwell equations of the electromagnetic field are

$$\square A_\mu = 4\pi j_\mu, \quad \frac{\partial A_\mu}{\partial x_\mu} = 0, \quad (1)$$

where A_μ is the potential 4-vector and j_μ represents the charge-density and current 4-vector. The field tensor $F_{\mu\nu}$ is defined in terms of the potentials by the relation

$$F_{\mu\nu} = \frac{\partial A_\nu}{\partial x^\mu} - \frac{\partial A_\mu}{\partial x^\nu}. \quad (2)$$

Let the co-ordinates of an electron be denoted by a 4-vector z_μ (sometimes written merely as z); z_μ is a function of the proper time τ of the electron. The current vector $j_\mu(x)$ at any space-time point x is given by

$$j_\mu(x) = e \int \dot{z}_\mu(\tau) \delta^4(x - z) d\tau; \quad (3)$$

$$\delta^4(x - z) = \delta(x_0 - z_0) \delta(x_1 - z_1) \delta(x_2 - z_2) \delta(x_3 - z_3).$$

Dot denotes differentiation with respect to the proper time τ . The retarded and the advanced proper times associated with any world point x are defined by

$$[x - z(\tau), x - z(\tau)] = 0, \quad (4)$$

where $(x_0 - z_0) > 0$ for the retarded proper time and $(x_0 - z_0) < 0$ for the advanced proper time.

We first consider the method of Riesz which has been applied to the classical fields by Fremberg and its quantization discussed by Gustafson. We begin by defining $A_\mu^\alpha(x)$ as (Fremberg 1946)

$$A_\mu^\alpha(x) = \frac{4\pi}{H(\alpha)} \int j_\mu(x') r^{\alpha-2} d^4x', \quad (5)$$

where
$$H(\alpha) = 2^{\alpha+1} \pi \Gamma\left(\frac{\alpha}{2}\right) \Gamma\left(\frac{\alpha}{2} + 1\right),$$

and
$$r^2 = [x - x', x - x'].$$

The domain of integration consists of the four-dimensional region bounded by the retrograde light cone having its vertex at x and the three-dimensional hyperplane

$x'_0 = -\infty$. $j_\mu(x)$ describes the source distribution: $\frac{\partial j_\mu(x)}{\partial x_\mu} = 0$. It is readily established that

$$\frac{\partial A_\mu^\alpha(x)}{\partial x_\mu} = 0, \quad \square A_\mu^\alpha(x) = A_\mu^{\alpha-2}(x). \quad (6)$$

The integral defining $A_\mu^\alpha(x)$ is convergent for $\alpha > 0$. By analytic continuation $A_\mu^\alpha(x)$ can be defined for all values of α in the α -plane. It can be shown that

$$A_\mu^{-2}(x) = 4\pi j_\mu(x), \quad (7)$$

where $A_\mu^{-2}(x)$ stands for the analytic continuation to $\alpha = -2$ of $A_\mu^\alpha(x)$, and, therefore, $A_\mu^\alpha(x)$ analytically continued to $\alpha = 0$ represents the solution of Maxwell equations (1). In the case of a point charge describing the world line $z_\mu(\tau)$, (5) reduces to

$$A_\mu^\alpha(x) = \frac{4\pi e}{H(\alpha)} \int_{-\infty}^{\tau_-} \dot{z}_\mu(\tau) r^{\alpha-2} d\tau, \quad (8)$$

where τ_- is the retarded proper time associated with x . [To carry out the analytic continuation we note the general theorem (see, for instance, Ma (1947)): for any function $f(\xi)$ which can be expanded into a Laurent's series of the form

$$f(\xi) = \sum_{-m}^{\infty} a_n \xi^n \quad (m \geq 0),$$

when analytic continuation is performed to $\alpha = 0$, we have

$$\alpha \int_0^X f(\xi) \xi^{\alpha-1} d\xi \rightarrow a_0, \quad (9)$$

whatever X may be. a_0 is the 'finite part' of $f(\xi)$.]

When x does not lie on the world line, $A_\mu(x)$ gives the Maxwell retarded potential for a point charge, and when x lies on the world line we obtain (Ma 1947) from (8),

$$A_\mu(z) = \frac{1}{2}[A_{\mu-}(z) - A_{\mu+}(z)], \quad (10)$$

where the suffixes $+$ and $-$ indicate that in the expressions τ refers to the advanced proper time and the retarded proper time respectively.

As an illustration of the method we consider the case of an electron at rest. Noting that for $\alpha \rightarrow 0$, $H(\alpha)$ reduces to $4\pi/\alpha$, (8) gives

$$A_0^\alpha(x) = e\alpha \int_{-\infty}^{\tau_-} r^{\alpha-2} d\tau, \quad A_l^\alpha(x) = 0 \quad (l = 1, 2, 3).$$

Again

$$r dr = -[x - z(\tau), \dot{z}(\tau)] d\tau = -\sqrt{(r^2 + |r|^2)} d\tau,$$

where $|r|$ is the length of the three-dimensional part of r . Hence we have

$$A_0^\alpha(x) = e\alpha \int_0^\infty \frac{r^{\alpha-1}}{\sqrt{(r^2 + |r|^2)}} dr, \quad (11)$$

which on using (9) immediately gives

$$A_0(x) = \frac{e}{|r|} \quad \text{for } |r| > 0,$$

$$= 0 \quad \text{for } |r| = 0.$$

and

(The analytic continuation of $A_\mu^\alpha(x)$ for $\alpha = 0$ is written simply by $A_\mu(x)$.) Thus at every accessible point the Riesz potential for an electron at rest reduces to the Coulomb value, but at the electron itself its value sinks to zero. The divergence at $|r| = 0$ is automatically eliminated. In the case of the field we find by first differentiating (11) with respect to $|r|$ and then performing analytic continuation that it has the Coulomb value everywhere except at $|r| = 0$, where it vanishes. Similarly, the energy of the electrostatic field outside a sphere of finite radius $|r|$ is $e^2/2|r|$, but the total energy of the field (the electrostatic self-energy) instead of diverging has the value zero. It should be noted that the self-energy vanishes in spite of the fact that the field and the energy density have the Coulomb value at all points (except $|r| = 0$).*

The Riesz potential thus gives the retarded potential for all points not on the world line, and on the world line it reduces to half the difference of the retarded and the advanced potentials. Thus, we see that the analytic continuation of (8) to $\alpha = 0$ gives the same result as the Dirac-Wentzel method. Ma (1947) has also established the equivalence of the two methods for a general point source.

3. We now turn to the method of residues (Frenkel 1926). For a point source, describing the world line $z_\mu(\tau)$, consider in the τ -plane the contour integral

$$I = -\frac{1}{i\pi} \int_C \frac{S(\tau)}{s^2} d\tau; \quad s^2 = [x - z(\tau), x - z(\tau)]. \quad (12)$$

s^2 vanishes for $\tau = \tau_-$ (retarded proper time) and for $\tau = \tau_+$ (advanced proper time). (In (12) τ is a complex quantity; when τ is real we write τ^2 , as in § 2, in place of s^2 .) For the contour C we have the following four cases:

- (i) C does not include τ_- and τ_+ : $I = 0$.
- (ii) C includes τ_- but not τ_+ : I gives the retarded potential.
- (iii) C includes τ_+ but not τ_- : I gives the advanced potential (with a negative sign).
- (iv) C includes both τ_- and τ_+ : I gives the difference of the retarded and the advanced potentials.

We therefore define $A_\mu(x)$ by

$$A_\mu(x) = -q \frac{e}{i\pi} \int_C \frac{\dot{z}_\mu(\tau)}{s^2} d\tau, \quad (13)$$

* It is interesting to compare this result with that of Pryce (1938). The energy-momentum tensor defined by Pryce makes the energy density for an electron at rest vanish throughout space, and the self-energy vanishes as a consequence of it.

where $q = 1$ for $x \neq z$, and $q = \frac{1}{2}$ for $x = z$. The contour C is to be taken about the retarded point only for $x \neq z$; when $x = z$ the retarded and the advanced points coincide and C includes, therefore, the advanced point as well. It is readily shown that when $x \neq z$, (13) gives the retarded potential, and for $x = z$ it reduces to half the difference of the retarded and the advanced potentials.

We shall show that equation (13) is equivalent to (8) of Riesz's method. Consider the case when x lies on the world line ($x = z(\tau_0)$). We split the range of integration in (8) from $-\infty$ to $(\tau_0 - \epsilon)$ and from $(\tau_0 - \epsilon)$ to τ_0 (ϵ is a small quantity). The integral from $-\infty$ to $(\tau_0 - \epsilon)$ is convergent for $\alpha < 1$, and, therefore, it vanishes for $\alpha = 0$. We expand the integrand in the second integral in a power series of $\sigma = (\tau_0 - \tau)$, and we obtain

$$r^2 = \sigma^2 - \frac{1}{12}\dot{v}^2\sigma^4 - \frac{1}{12}[\ddot{v}, \ddot{v}]\sigma^6 + O(\sigma^8), \quad (14)$$

and hence we have

$$\begin{aligned} A_\mu^\alpha(x) &= \frac{4\pi e}{H(\alpha)} \int_{\tau_0-\epsilon}^{\tau_0} \dot{z}_\mu(\tau) r^{\alpha-2} d\tau \\ &\sim e\alpha \int_0^\epsilon \dot{z}_\mu(\tau_0 - \sigma) \left\{ 1 - \frac{1}{12}\dot{v}^2\sigma^2 + O(\sigma^4) \right\}^{\frac{1}{2}\alpha-1} \sigma^{\alpha-2} d\sigma. \end{aligned} \quad (15)$$

When analytic continuation to $\alpha = 0$ is performed, (15), according to (9), reduces to the 'finite part' of the expression

$$\frac{e\dot{z}_\mu(\tau_0 - \sigma)}{\sigma\{1 - \frac{1}{12}\dot{v}^2\sigma^2 + O(\sigma^4)\}} = \frac{e\dot{z}_\mu(\tau_0 - \sigma)\sigma}{r^2} = -\frac{e\dot{z}_\mu(\tau)(\tau - \tau_0)}{r^2},$$

which in terms of a contour integral is

$$-\frac{1}{2i\pi} \int_C \frac{e\dot{z}_\mu(\tau)}{s^2} d\tau,$$

the contour C being round the point $\tau = \tau_0$. For the case when x does not lie on the world line, (14) gets replaced by (in the neighbourhood of the retarded point τ)

$$r^2 = -2\kappa(\tau_-)(\tau - \tau_-) - (1 - \kappa'(\tau_-))(\tau - \tau_-)^2 - \frac{1}{3}\kappa''(\tau_-)(\tau - \tau_-)^3 + O((\tau - \tau_-)^4), \quad (16)$$

where

$$\kappa(\tau) = [\dot{z}, x - z],$$

$$\kappa'(\tau) = [\ddot{z}, x - z],$$

$$\kappa''(\tau) = [\ddot{\ddot{z}}, x - z].$$

Proceeding as before it can be shown that on analytic continuation to $\alpha = 0$, (8) reduces to (13). The equivalence for a general point source for an electromagnetic as well as meson field can be similarly established.

REFERENCES

- Fremberg, N. F. 1946 *Proc. Roy. Soc. A*, **188**, 18.
 Frenkel, J. 1926 *Lehrbuch der Elektrodynamik*, p. 177.
 Ma, S. T. 1947 *Phys. Rev.* **71**, 787.
 Pryce, M. H. L. 1938 *Proc. Roy. Soc. A*, **168**, 389.

Hot-wire investigation of the wake behind cylinders at low Reynolds numbers .

By L. S. G. KOVÁSZNAY

*Department of Aeronautics, The Johns Hopkins University,
Baltimore, Md., U.S.A.*

(Communicated by Sir Geoffrey Taylor, F.R.S.—Received 17 December 1948)

The hot-wire technique has been used to measure the regular vortex street pattern behind a cylinder at low Reynolds number. Measurements of mean velocity distribution were made both below and above the critical Reynolds number at which the periodic motion appears.

Amplitude and phase measurements gave sufficient information for computation of the instantaneous flow pattern of the vortex system. The important points resulting from the investigation are that (i) the critical Reynolds number at which vortices are shed is 40, (ii) in the range of Reynolds numbers investigated the vortices are not shed directly from the cylinder but appear some distance downstream as an instability of the laminar wake.

INTRODUCTION

The regular pattern of the vortex street behind cylinders at low Reynolds numbers has attracted considerable attention in the past (Kármán 1912). The theoretical treatment of the problem, since it is based on the assumption of a perfect fluid, does not account for the Reynolds number effect. The time frequency of the vortex pattern varies with Reynolds number, and this variation has been measured by several authors (Fage & Johansen 1927; Relf & Ower 1923; Richardson 1923; Rosenhead & Schwabe 1930).

The sudden appearance of the periodic wake at a certain Reynolds number with a definite pattern indicates a laminar instability phenomenon.

The hot-wire anemometer seemed to be a suitable tool for detailed measurements at the low speeds and small linear dimensions involved.

The work was started at the Department of Aerodynamics, Royal Hungarian University for Technical Sciences, in 1943; it was interrupted during the war, and continued in the Cavendish Laboratory, Cambridge, England.

EQUIPMENT

The equipment used in Budapest and that used in Cambridge were quite similar:

	Budapest (1943)	Cambridge (1946)
wind tunnel working section	20 by 20 in.	15 by 15 in.
speed range	1.5 to 80 ft./sec.	1 to 60 ft./sec.
background turbulence level	0.18 %	0.06 %
hot wires (platinum)	0.0002 in. diam.	0.0001 in. diam.
hot-wire equipment	see Kovasznyay (1947)	see Townsend (1947)

Measurement of the velocity of the undisturbed flow was made in Budapest by Pitot head and high-sensitivity Reichart type torsion balance micromanometer. This gave an accuracy of 1 % down to velocities as low as 50 cm./sec.

In Cambridge no micromanometer of this low range was available, so a special low-speed meter was built.

The need for measuring very low mean speed first suggested the use of hot-wire technique. But since the calibration of a thin hot wire is reliable for only a limited interval of time, a more rugged instrument was desirable. The instrument finally built also worked on the cooling effect of the air stream.

A steel hypodermic needle (0.7 mm. outside diameter) was heated by 5 to 10 amp. a.c., from the secondary windings of a special transformer. A copper-constantan thermocouple was placed in a glass capillary inside the hypodermic needle.

Two identical units were made, and the balanced output of the two thermocouples, oppositely connected, was fed into a millivoltmeter (Weston, 5 mV range, 5Ω resistance). One unit was placed in the air stream, the other was kept at constant mean temperature of the room, and served as reference for maintaining constant heating current.

The instrument was calibrated in the wind tunnel in the following way: a thin (0.060 in.) resistance wire was mounted across the wind tunnel and heated by d.c. with a superimposed a.c. (50 c./sec.), resulting in a periodic temperature fluctuation in the wake. The heat 'patches' could be identified in the laminar wake of the wire quite well by comparing the signal of an exploring hot-wire anemometer with this a.c. heating voltage, in particular, by noting the phase relation. From the wavelength of the heat pattern in the wake, the mean speed was determined. (The instrument was usable in the speed range of 1 to 10 ft./sec.)

MEASUREMENT OF FREQUENCY AND OF CRITICAL REYNOLDS NUMBER

The frequency of the periodic wake was observed with the hot-wire anemometer. The maximum signal was obtained when the hot wire was located 5 to 10 diameters behind the wire and slightly off the centre of the wake.

The hot-wire signal was amplified and fed into the horizontal deflexion system of a cathode-ray oscilloscope; an audio-frequency oscillator was used for the vertical deflexion. Thus when the two frequencies were equal, the Lissajous figure became an ellipse.

This method gave sufficient accuracy at relatively low Reynolds numbers where the motion was extremely regular; but at higher Reynolds numbers, decreasing regularity reduced the accuracy.

The results are expressed in the form of Strouhal number:

$$S(R) = \frac{fd}{U_0}, \text{ where } R = \frac{U_0 d}{\nu},$$

where f is the observed frequency, d the diameter of the wire, U_0 the undisturbed mean speed, ν kinematic viscosity. Figure 1 is a logarithmic plot of the results using wires of various diameters.

Careful measurements were made to determine the Reynolds number at which the periodic wake appears. The critical Reynolds number was found to be $R_{\text{crit.}} = 40$. This value was obtained with wires of very high length/diameter ratio (the length

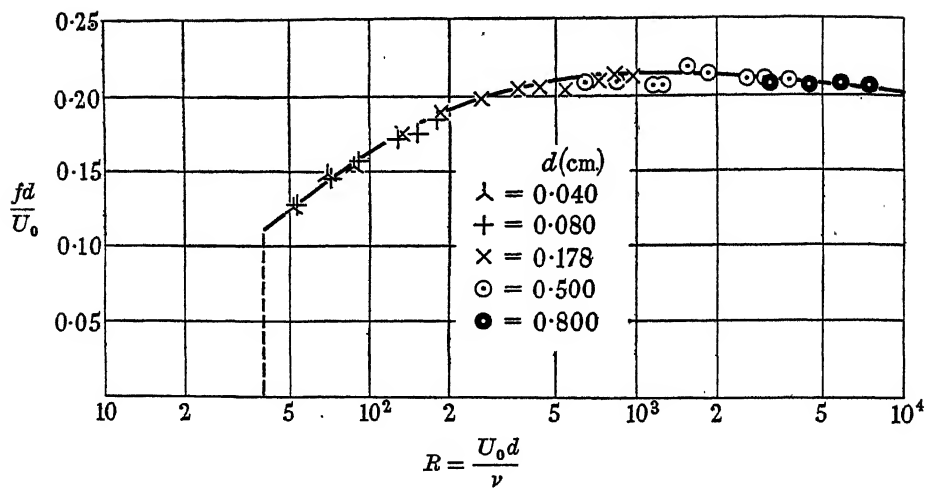


FIGURE 1

TABLE 1. MEAN VELOCITY DISTRIBUTION IN THE LAMINAR WAKE, $R = 34$

$x = 2d$		$x = 5d$		$x = 6d$	
y/d	U/U_0	y/d	U/U_0	y/d	U/U_0
-4.76	1.018	-3.95	1.035	-3.40	1.012
-4.13	1.037	-2.65	1.053	-2.76	1.012
-3.65	1.045	-1.98	1.035	-2.17	1.012
-2.87	1.073	-1.40	0.910	-1.50	0.857
-2.25	1.110	-1.07	0.687	-0.81	0.468
-1.65	1.155	-0.76	0.425	-0.50	0.280
-1.26	1.090	-0.42	0.216	0	0.180
-0.98	0.823	0	0.153	+0.29	0.243
-0.81	0.550	+0.26	0.193	+0.78	0.506
-0.64	0.300	+0.58	0.348	+1.18	0.784
-0.33	0.094	+0.83	0.538	+1.79	0.950
-0.02	0.065	+1.12	0.738	+2.41	0.988
+0.32	0.102	+1.46	0.884	+3.02	0.988
+0.65	0.326	+2.11	0.960		
+0.81	0.578	+3.34	0.973		
+0.99	0.833				
+1.50	1.018				
+1.93	1.095				
+4.88	0.990				

$x = 10d$		$x = 20d$	
y/d	U/U_0	y/d	U/U_0
-3.79	1.000	-3.79	1.000
-2.55	0.946	-3.15	1.000
-1.90	0.915	-2.54	0.891
-1.27	0.742	-1.89	0.861
-0.81	0.596	-1.26	0.755
-0.48	0.454	-0.65	0.630
+0.10	0.355	+0.02	0.596
+0.53	0.482	+0.65	0.630
+0.93	0.672	+1.30	0.865
+1.26	0.808	+1.96	0.983
+1.89	0.974	+2.56	1.015
+2.54	1.035		
+3.85	1.005		

of the wire was always 50 cm.). The transition does not seem to have any hysteresis; the critical value is the same for both increasing and decreasing mean velocity.

If the velocity is set very slightly below the critical value a slight disturbance, e.g. 'plucking' the wire, started the development of periodicity, but it decayed

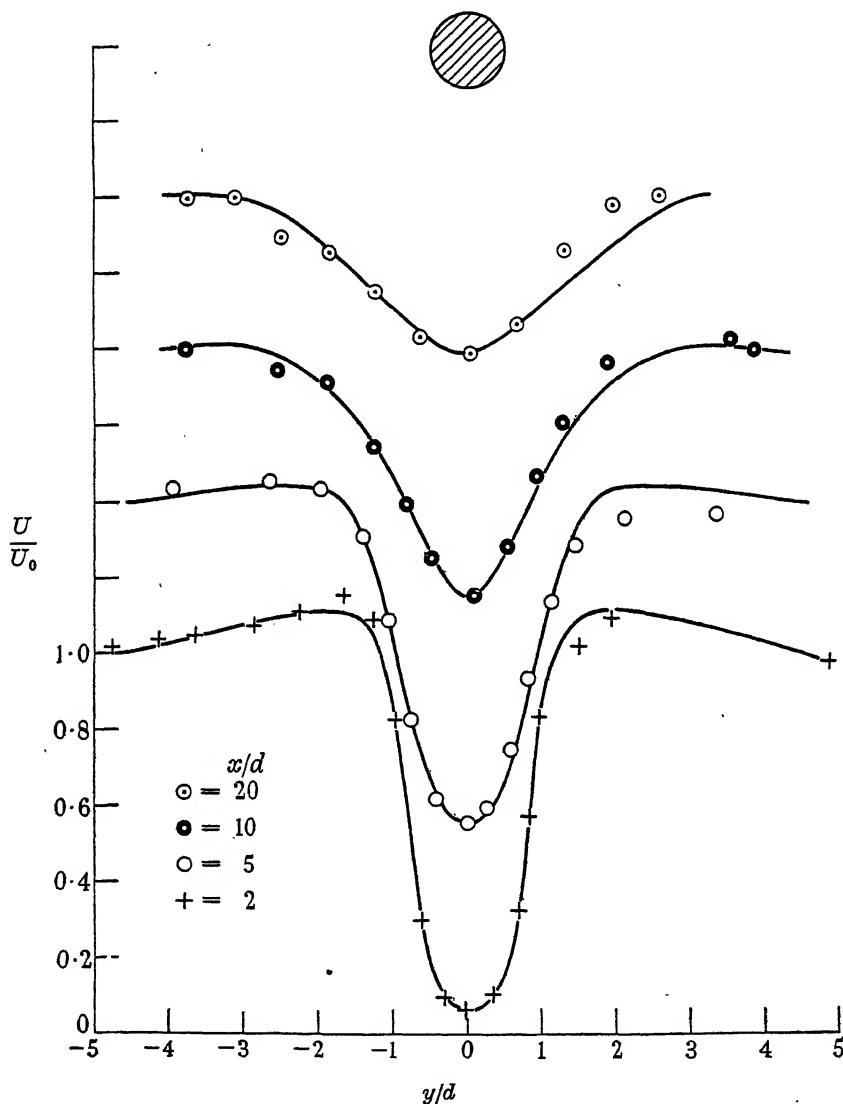


FIGURE 2

slowly as the external disturbance ceased. This phenomenon gives strong evidence for the existence of laminar oscillations in the viscous wake at a certain critical Reynolds number. In earlier measurements the wall effects seemed to modify the results (Thom 1933).

TABLE 2. MEAN VELOCITY AND R.M.S. VELOCITY FLUCTUATION
IN THE PERIODIC WAKE, $R = 56$

$x = 2d$			$x = 3.5d$			$x = 5d$		
y/d	U/U_0	$100 \frac{u'}{U_0}$	y/d	U/U_0	$100 \frac{u'}{U_0}$	y/d	U/U_0	$100 \frac{u'}{U_0}$
-3.64	1.000	0.17	-5.23	1.000	0.26	-5.04	1.000	0.43
-3.02	1.045	0.38	-2.82	1.070	1.01	-3.64	1.010	0.76
-1.73	1.120	0.72	-1.93	1.080	2.47	-3.04	1.020	1.34
-1.41	1.120	0.88	-1.16	0.930	5.78	-2.39	1.020	2.63
-1.07	1.000	1.02	-0.84	0.660	8.70	-1.78	1.025	5.00
-0.86	0.773	1.06	-0.53	0.289	6.97	-1.14	0.880	9.11
-0.51	0.152	0.42	-0.22	0.080	1.94	-0.79	0.639	12.18
-0.24	0.048	0.12	-0.10	0.068	1.13	-0.51	0.472	12.05
0	0.039	0.07	0	0.062	0.70	-0.39	0.403	9.95
+0.25	0.048	0.15	+0.13	0.068	1.09	-0.19	0.300	5.96
						-0.09	0.305	5.58
						0	0.281	3.96
						+0.11	0.276	4.88
						+0.25	0.291	7.65
						+0.25	0.351	7.63
						+0.42	0.443	10.63

$x = 8d$			$x = 12d$			$x = 20d$		
y/d	U/U_0	$100 \frac{u'}{U_0}$	y/d	U/U_0	$100 \frac{u'}{U_0}$	y/d	U/U_0	$100 \frac{u'}{U_0}$
-6.36	1.000	0.09	-10.32	1.00	0.19	-13.30	1.00	0.24
-5.65	1.020	0.32	-5.72	1.00	0.40	-6.35	1.00	0.32
-4.35	1.040	0.95	-5.08	1.00	0.74	-5.21	1.00	0.91
-3.73	1.030	1.64	-4.45	1.00	1.20	-4.57	1.00	1.46
-3.10	1.030	2.87	-3.81	0.99	2.16	-3.94	1.00	2.52
-2.50	1.020	5.38	-3.18	0.99	3.95	-3.30	1.00	4.24
-1.84	0.990	8.70	-2.54	0.97	6.66	-2.67	0.96	5.60
-1.54	0.950	11.32	-1.90	0.92	9.10	-2.03	0.92	5.96
-1.23	0.870	13.50	-1.46	0.85	9.55	-1.40	0.86	—
-0.85	0.805	13.38	-0.83	0.77	7.40	-0.94	0.81	3.51
-0.47	0.635	9.13	-0.51	0.74	6.04	0	0.79	2.30
0	0.590	5.18	-0.29	0.73	5.00	+1.11	0.85	4.61
+0.51	0.647	9.32	0	0.69	4.25			
+0.84	0.720	11.85	0	0.69	4.04			
+1.12	0.800	11.95	+0.33	0.72	8.42			
+1.66	0.945	9.40	+1.41	0.85	8.42			
+2.26	0.980	5.62						
+3.19	0.980	2.14						
+4.53	0.980	0.59						

$x = 40d$			$x = 100d$		
y/d	U/U_0	$100 \frac{u'}{U_0}$	y/d	U/U_0	$100 \frac{u'}{U_0}$
-11.6	1.00	0.14	-4.07	1.00	0.43
-6.3	1.00	0.37	-2.54	0.96	0.90
-5.4	1.00	1.00	-1.91	0.91	0.88
-4.12	1.00	2.30	-1.27	0.91	0.59
-2.86	0.96	3.35	-0.64	0.87	0.38
-2.22	0.89	2.63	0	0.83	0
-1.59	0.86	1.86	+2.03	0.88	0.76
-0.95	0.83	1.27			
0	0.80	1.00			

MEASUREMENTS OF MEAN FLOW

Experiments were carried out to obtain the velocity distribution behind a cylinder both slightly below and slightly above the critical Reynolds numbers. The cylinder used was a 1 mm. steel wire. The measurement was difficult because of the low speed and the necessarily small traversing steps.

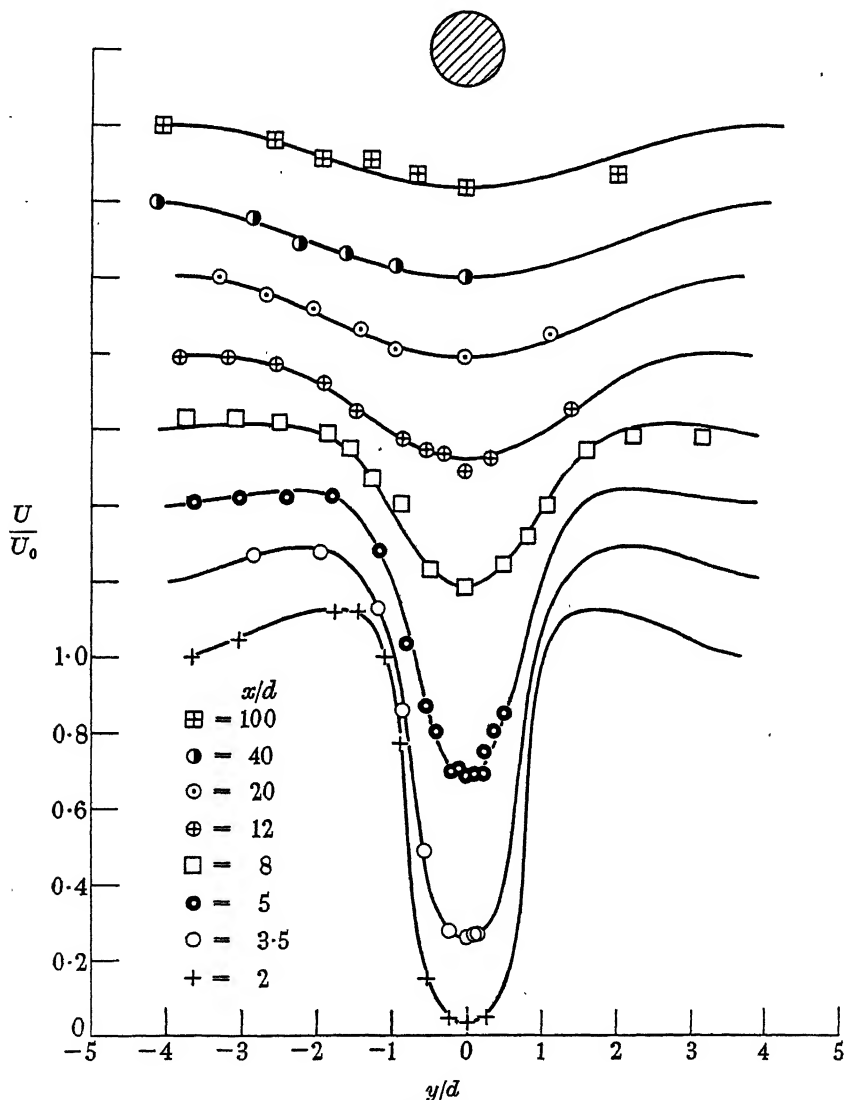
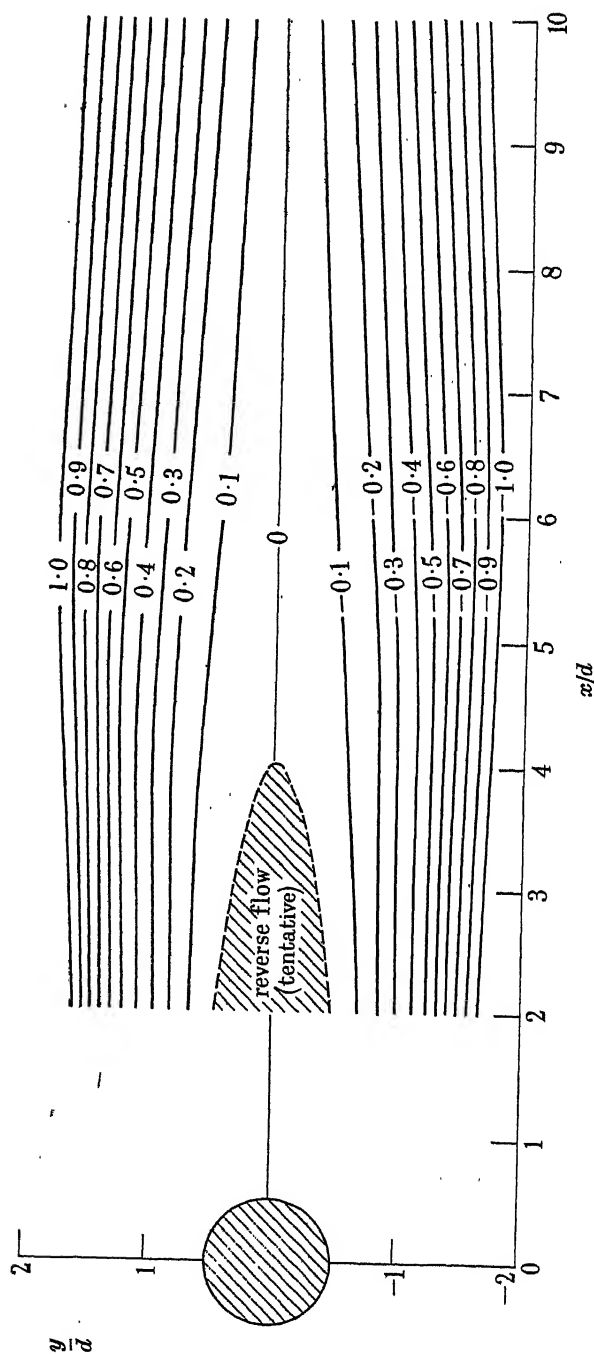


FIGURE 3

The hot wire should not be supported by long prongs parallel to the flow, because this caused an instability in the wake that propagated upstream. Therefore the probe had to enter the wake from the side. This caused slight asymmetry in the reading, but it was assumed that the half-distribution measured before the centre was reached was unaffected. The data are given in table 1. Figure 2 shows the velocity distribution for the laminar case at $R = 34$.

FIGURE 4 ($R = 34$)

The measurement of the mean velocity in the periodic wake is more difficult. As the velocity fluctuates, the mean readings of the hot wire are affected by the non-linear response, so that a 'rectifying effect' appears. Correction of the mean velocity readings was made on the basis of non-linear hot-wire characteristics with the use of the velocity fluctuation measurements at the same point. The treatment is identical with the computation of the distortion of a sine wave due to non-linear distortion of an amplifier stage.

The fluctuations made detection of zero mean velocity almost impossible, since a hot-wire set parallel to the z -axis measures absolute value of the velocity vector in a two-dimensional (x, y) flow. The data of mean speed and fluctuation are given in table 2.

Figure 3 shows the mean velocity distribution behind the 1 mm. diameter wire at Reynolds number $R = 56$. The vortex street is fully developed.

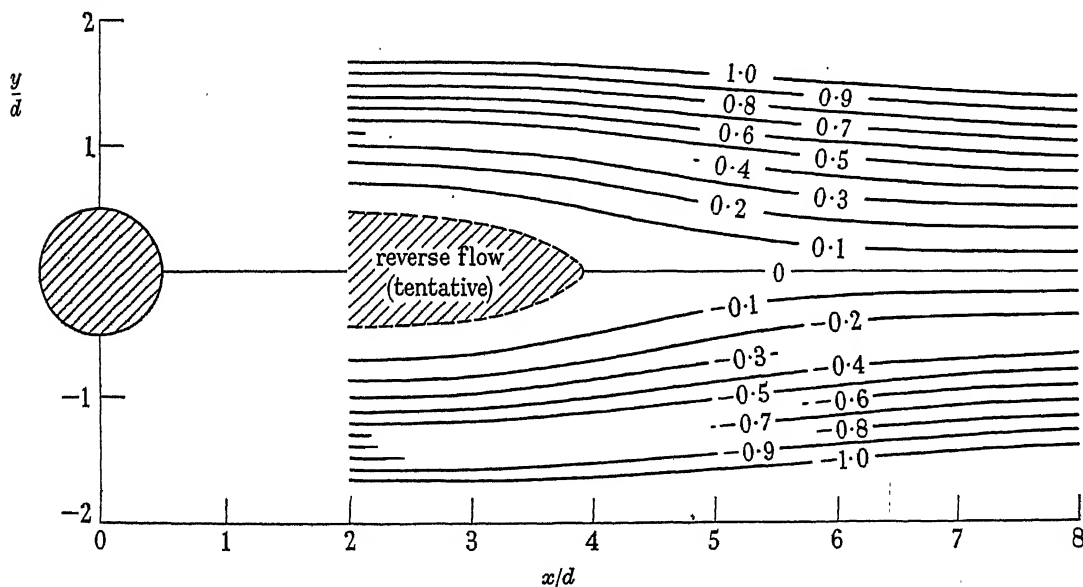


FIGURE 5 ($R = 56$)

MEAN FLOW PATTERN

The two-dimensionality of the flow was experimentally verified. Thus the hot-wire measurement gives the absolute magnitude of the velocity, and therefore a stream function ψ_0 exists:

$$U_x = \frac{\partial \psi_0}{\partial y}, \quad U_y = -\frac{\partial \psi_0}{\partial x}, \quad U = \sqrt{(U_x^2 + U_y^2)} = \left| \frac{\partial \psi_0}{\partial n} \right|,$$

where U_x , U_y are the components of mean velocity and n is the direction perpendicular to the stream lines. If $U(x, y)$ is known from measurement, the ' $\psi_0 = \text{constant}$ ' lines (streamlines) can be constructed graphically by using small circles with diameter $1/U$, and drawing the ' $\psi_0 = \text{constant}$ ' lines tangential to the circles.

The data given in figures 2 and 3 were used to construct figures 4 and 5.

Of course reversed flow cannot be detected with the standard type of single hot-wire anemometer. The scale of the flow phenomenon was so small that the use of direction meters was impracticable. Therefore it can only be stated that the shaded areas are probably regions of reversed flow. In these regions no substantial change in stream function could be detected.

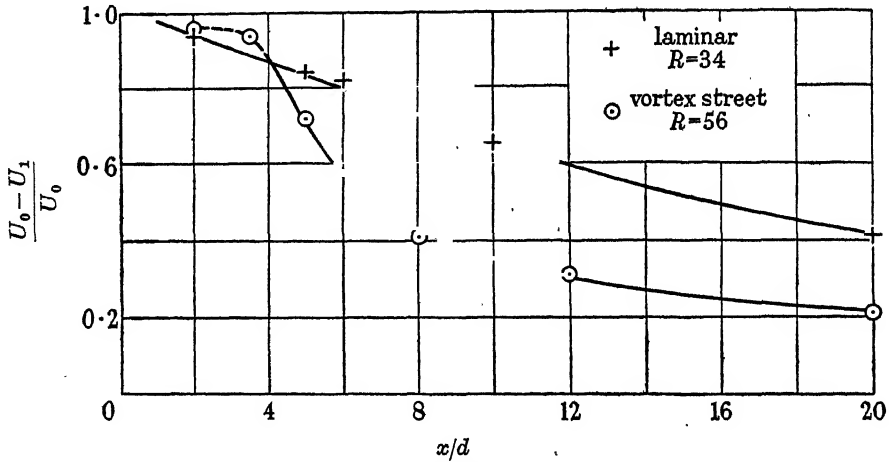


FIGURE 6

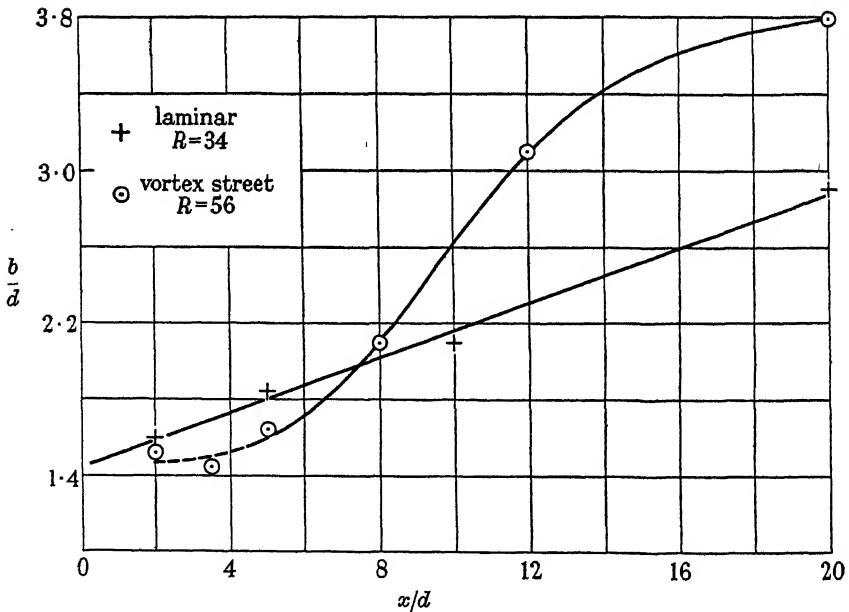


FIGURE 7

The main difference between the mean flows below (figure 4) and above (figure 5) the critical Reynolds number is that in the simple laminar flow the wake is more extensive. In the vortex street, the higher shearing stresses even out the differences more rapidly. In these two figures the spacing of the contour lines corresponds to

increments in the stream function, $\Delta\psi_0 = 0.1dU_0$, and the numbers are values of ψ_0/dU_0 .

Figure 6 shows the variation of the velocity along the centre of the wake ($U_1 = U$, where $y = 0$), and figure 7 shows the variation of the 'half-width' of the wake, b , defined by $y = \frac{1}{2}b$, where $U = \frac{1}{2}(U_0 + U_1)$.

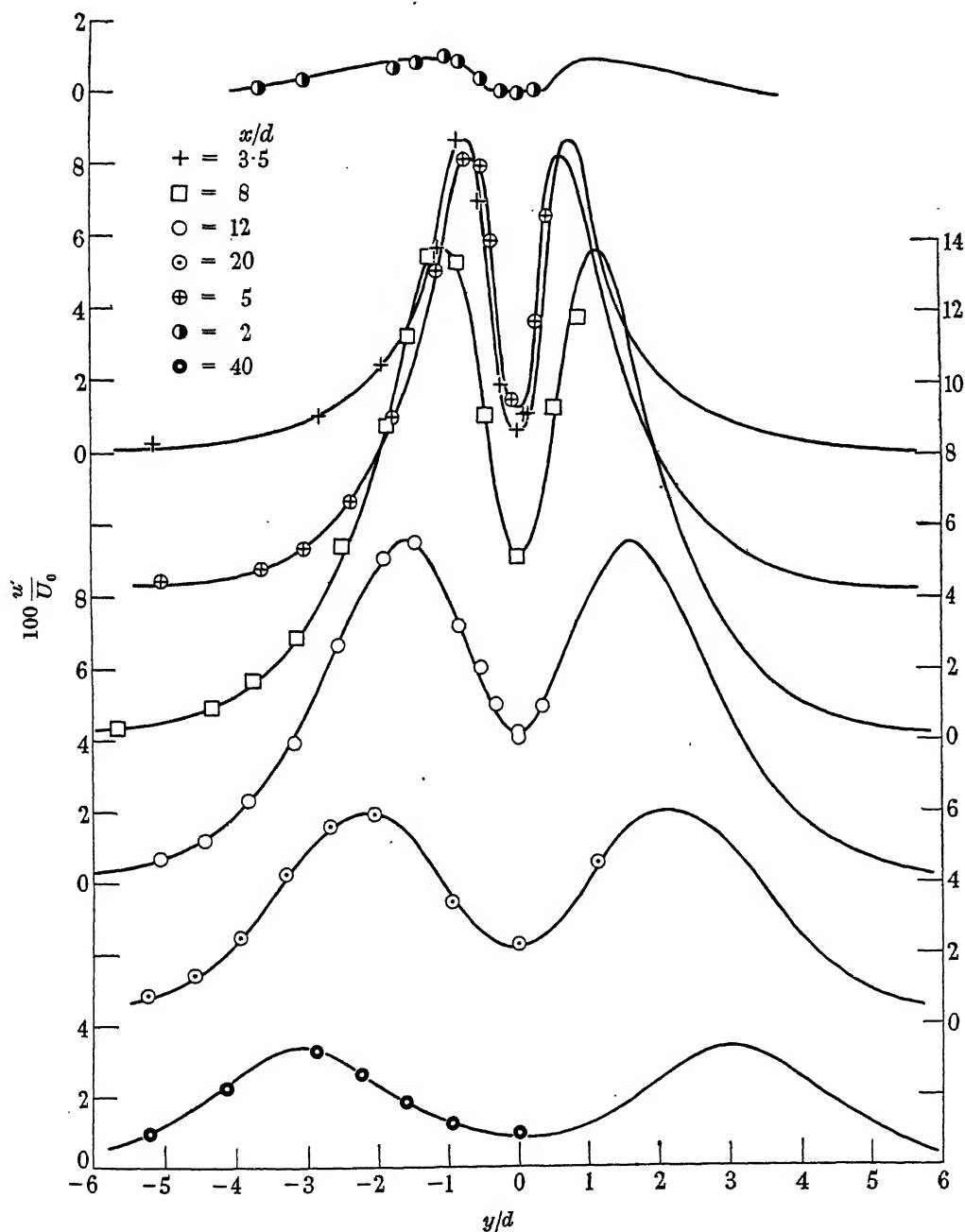
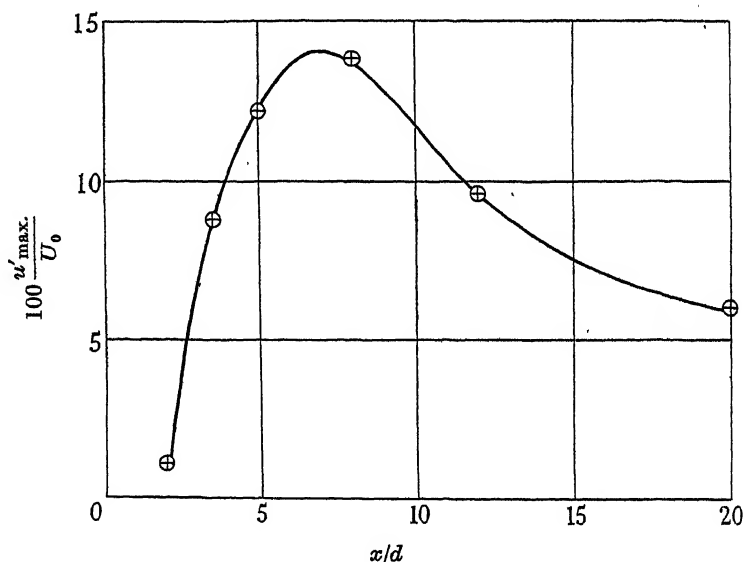


FIGURE 8

MEASUREMENT OF 'ROOT-MEAN-SQUARE' FLUCTUATIONS

Since the velocity fluctuations in the vortex street were almost one order of magnitude smaller than the mean velocities it was possible to determine the amplitude of this periodic motion by a very convenient method; the root-mean-square velocity fluctuation, as determined by a simple hot-wire ' u '-meter gave a good measure of this amplitude. Just as in the case of measurement of low turbulence levels, the second order contribution from the v' component of velocity fluctuation can be neglected.

The data for the mean flow and the fluctuations are given in tables 1 and 2. It is remarkable that the fluctuations close to the cylinder are very small and that they develop a maximum of intensity only in the vicinity of $x = 7d$. The distribution of fluctuation in traverses at $x/d = 2, 3.5, 5, 8, 12, 20$, are given in figure 8. The individual diagrams are staggered, and the ordinate scales alternate between left and right sides of the figure.

FIGURE 9 ($R = 56$)

The closest traverse, at $x/d = 2$, shows that very near to the cylinder there is practically no fluctuation present. Thus, the vortices are not 'shed' from the cylinder at this low Reynolds number, but develop several diameters downstream. Measurements were made on one side of the wake only, because, as mentioned earlier the hot-wire holder interfered with the wake when it crossed the middle portion. The oscillograph records showed that the instantaneous pattern then became slightly unstable.

The maxima of the fluctuations are plotted against downstream distance in figure 9. The maximum value of 14 % occurs at $x/d = 7$.

In general the fluctuations studied here are of pure periodic type; the pattern remained unchanged on the oscilloscope screen for hours, showing a surprising

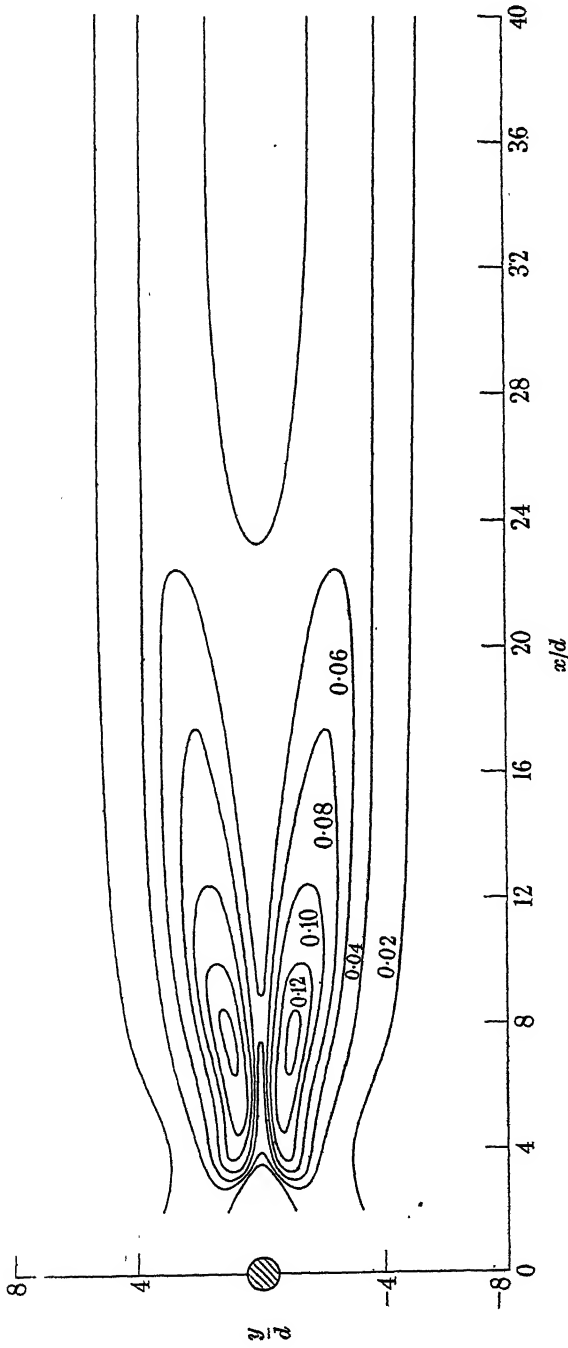


FIGURE 10 ($R = 56$)

stability. As was to be expected for such a flow pattern, the fluctuation at the middle of the wake was always pure double frequency. The double-frequency component decreased rapidly with distance from the wake centre, and disappeared almost completely before the point of maximum fluctuation was reached.

The contour diagram of the r.m.s. fluctuations was constructed in figure 10. The difference between contour lines is 2% of the undisturbed mean speed. ($\Delta u'/U_0 = 0.02$.)

INSTANTANEOUS-FLOW PATTERN

Since the flow pattern is completely regular and stable at Reynolds numbers not exceeding $R = 160$, phase measurements were possible so that the complete velocity distribution could be obtained.

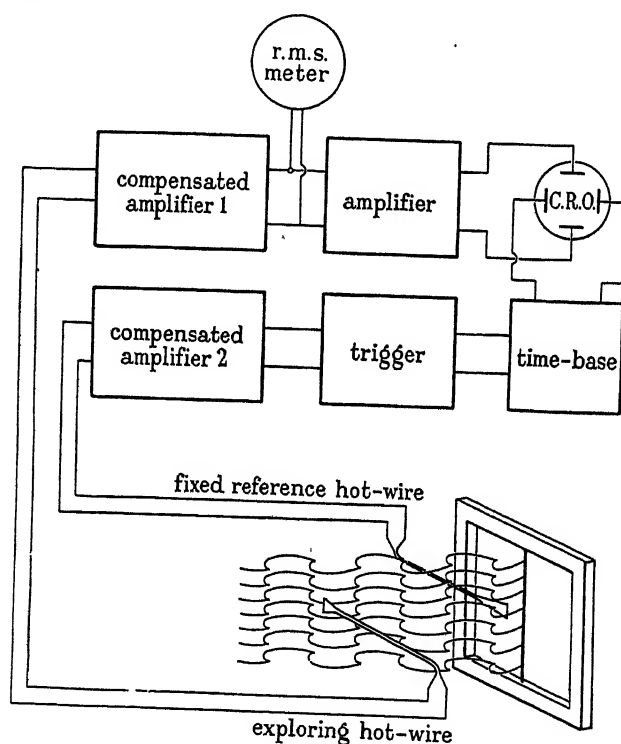


FIGURE 11

Figure 11 shows the block diagram of the electronic set-up used in point-to-point phase determination. One fixed hot wire served as reference for timing, and the other was mounted on a traversing mechanism. The reference hot wire was used to synchronize the time base (i.e. the sweep frequency) of the cathode-ray oscilloscope. This way a very steady and highly reproducible oscillogram could be obtained.

When the exploring hot wire was moved along, the pattern on the screen of the oscilloscope shifted correspondingly; therefore phase relations could be easily determined. The distance between points of the same phase angle gave the spacing or wave-length of the pattern. Typical oscillograph records are shown in figure 12.

For these records, $x/d = 5$, and the value of y/d is given on each record. The records were taken from synchronized steady patterns with exposures $\frac{1}{25}$ sec. Slight intensity variations are noticeable due to stroboscopic effect.

The relative phase of second and first harmonics varies with distance downstream. The first harmonic is asymmetric with respect to the x -axis, the second is symmetric. The phase relations are shown in table 3.

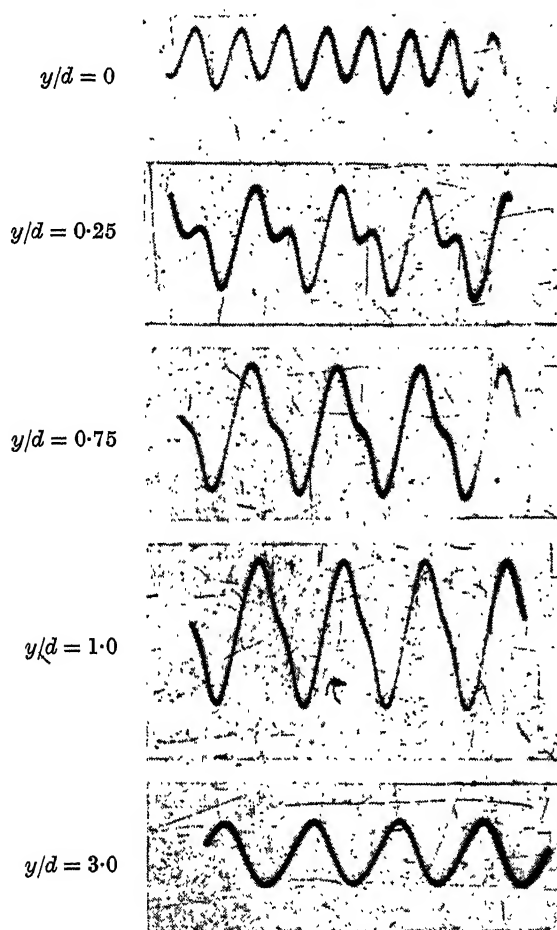


FIGURE 12

The velocity fluctuation distribution can be represented at any instant

$$U = U_1(x, y) \cos 2\pi[\zeta_1(x) + ft] + U_2(x, y) \cos 4\pi[\zeta_2(x) + ft],$$

where U_1 and U_2 vary slowly with x but rapidly with y . U_1 is the odd function of y ; U_2 is the even function. $\zeta_1(x)$ and $\zeta_2(x)$ are almost linear functions of x , taking care of non-uniform spacing. The time frequency f is constant because the sequence of eddies is continuous.

TABLE 3. PHASE OF THE FIRST AND SECOND COMPONENTS
OF THE PERIODIC WAKE AT $R = 56$

ζ_1	x/d	ζ_2	x/d
0	2	0	2
0.5	7.5	0.25	5.2
1.0	12.0	0.50	8.0
1.5	16.2	0.75	10.5
2.0	20.0	1.00	13.0
2.5	23.7	1.25	15.5
3.0	27.4	1.50	17.5
3.5	31.1	1.75	19.3
4.0	34.8	2.00	20.9
4.5	38.5	2.25	23.0
		2.50	25.1
		2.75	27.2
		3.00	29.5
		3.25	31.3
		3.50	33.0
		3.75	34.9
		4.00	36.7
		4.25	38.4

The instantaneous total stream function $\psi(x, y)$ was constructed by using the following simplifications.

(1) U_1 was computed as $\sqrt{2}u'$ in the region where no double frequency was visible.

(2) U_2 was computed on the assumption that it vanished parabolically in the $\pm y$ directions:

$$U_2(x, y) = U_2(x, 0) [1 - y^2/y_1^2],$$

where y_1 is the value of y for which the total fluctuation is a maximum.

(3) The velocity fluctuation measured by the hot wire was taken identical with velocity fluctuation in the x -direction.

The flow was constructed from the co-ordinate system fixed to mean motion of the fluid; the velocity is 0 at infinity. The computation of ψ was made in three steps: (a) the stream function of the mean flow (ψ_0), (b) the first harmonic of the fluctuation (ψ_1), (c) the second harmonic (ψ_2). The resulting pattern is given in figure 11, and the lines correspond to differences in the stream function, $\Delta\psi = 0.1U_0d$, the dotted lines are half-values between two full lines. The streamline pattern in figure 13 is viewed relative to the undisturbed flow at infinity. The diagram clearly shows the development and decay of this motion.

CONCLUSIONS

The experiments have established the following facts:

(1) The critical Reynolds number, i.e. that at which the vortex street sets in, is a well-defined value for infinitely long circular cylinders, and the pattern that appears is unique and stable.

(2) The 'vortices' develop some distance downstream within the Reynolds number range 40 to 160, and are not shed directly from the cylinder. Consequently the

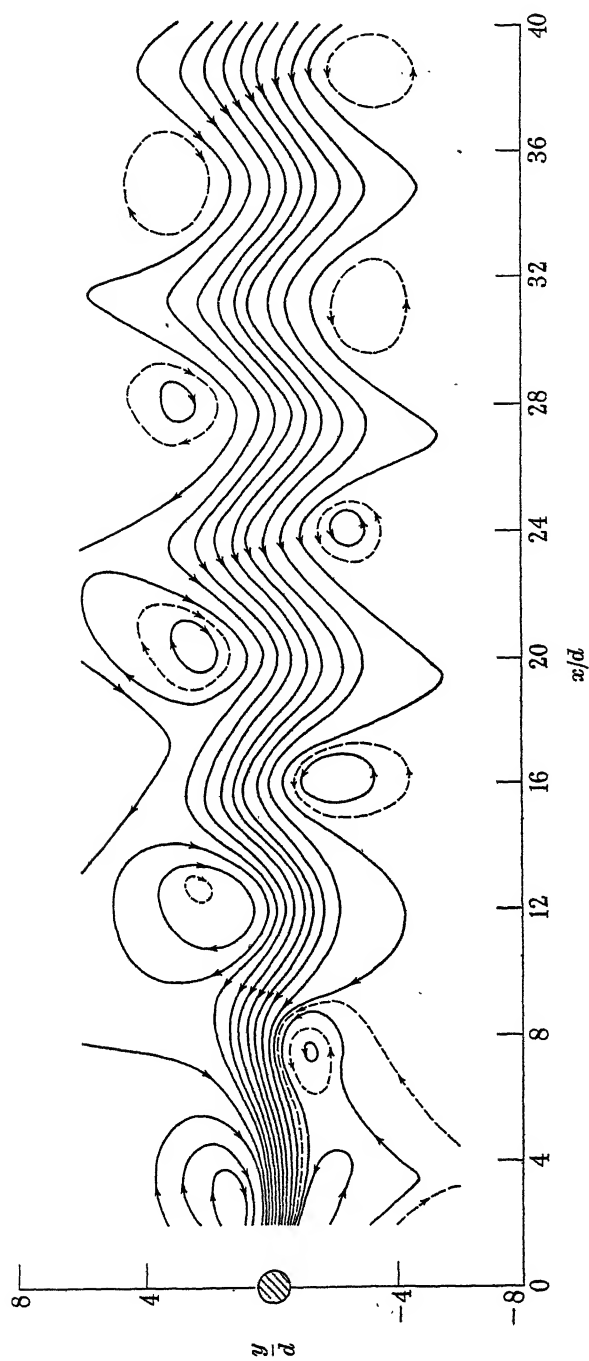


FIGURE 13 ($R = 56$)

phenomenon can properly be considered as an instability of the laminar wake, that develops up to an amplitude limit, but dies out before becoming turbulent, at least in this low Reynolds number range.

The major part of the present work was done in the Cavendish Laboratory, and the author wishes to express his indebtedness to Sir Geoffrey Taylor for his supervision and criticism. Thanks are due also to A. A. Townsend for making available his hot-wire apparatus.

The author was holding a British Council research scholarship when carrying out the research.

REFERENCES

- Fage, A. & Johansen, F. C. 1927 *Proc. Roy. Soc. A*, **116**, 170.
 Kármán, Th. v. 1912 *Nachr. Ges. Wiss. Göttingen*, p. 547.
 Kovásznay, L. S. G. 1947 *N.A.C.A. Tech. Mem.* no. 1130.
 Relf, E. F. & Ower, E. 1923 *A.R.C. Reports & Memoranda*, no. 825.
 Richardson, E. G. 1923 *Proc. Phys. Soc.* **36**, 153.
 Rosenhead, L. & Schwabe, M. 1930 *Proc. Roy. Soc. A*, **129**, 115.
 Thom, A. 1933 *Proc. Roy. Soc. A*, **141**, 651.
 Townsend, A. A. 1947 *Proc. Camb. Phil. Soc.* **43**, 560.

Plastic deformation of silver chloride

I. Internal stresses and the glide mechanism

By J. F. NYE, *Cavendish Laboratory, University of Cambridge*

(Communicated by E. Orowan, F.R.S.—Received 20 December 1948—

Revised 4 April 1949—Read 19 May 1949)

[Plates 5 to 8]

The experiments of Obreimow & Schubnikoff (1927) on the birefringence produced by the plastic deformation of single crystals of rock salt have been extended to a polycrystalline material. Rolled sheets of silver chloride have been recrystallized and then deformed plastically in various ways—by simple extension and by bending, for example. The sheets are transparent and very ductile and, since silver chloride is cubic in structure, the birefringence patterns observed under the microscope provide a picture of the distribution of the internal stresses uncomplicated by natural double refractions. It is suggested that results obtained with this optical method are applicable to metals.

Silver chloride appears to deform by glide, and when the glide packets are observed on edge a characteristic pattern of parallel birefringent bands is visible. The relation of the glide plane and glide direction to the crystal structure has been studied by making observations upon these bands and upon the glide lines formed on the surfaces of bars of square cross-section consisting effectively of chains of single crystals. The orientations of the fifteen sets of glide bands examined in this way were all consistent with glide movements in a $\langle 110 \rangle$ direction; the glide plane, however, was not always a crystallographic plane of low indices. In the six cases in which the measurement was possible, it lay within 9° of the plane in the $\langle 110 \rangle$ zone on which the maximum shear stress, resolved in the $\langle 110 \rangle$ direction, acted. It is concluded that silver chloride deforms by 'pencil glide', the mechanism postulated by Taylor & Elam in 1926 to explain the plastic behaviour of α -iron. The transmission of pencil glide across grain boundaries is discussed.

The residual stresses observed by the optical method in polycrystalline sheets may be divided into three groups: (1) A system of stresses set up between the glide zones of each grain and alternating with a period equal to the spacing of the glide zones. A detailed analysis of these is given in the second paper (part II). (2) Alternating stresses produced when a system of glide zones meets a grain boundary. (3) 'Heyn stresses' produced by the non-uniformity of plastic deformation from grain to grain.

1. INTRODUCTION

If one deforms a crystalline body plastically and removes the external forces it is found that stresses remain locked up in the interior. Internal stresses of this type may be present both in single crystals and in polycrystalline aggregates, and much attention has been paid in recent years to their occurrence in cold-worked, polycrystalline metals. The X-ray diffraction method, which has been the principal means of investigation, unfortunately suffers at present from a serious limitation: the stresses that are investigated are often distributed in space in patterns of microscopic scale, and they may vary considerably in distances comparable with the width of the X-ray beam. For this reason, an X-ray beam, while often capable of indicating the presence and the magnitude of internal stresses, is too coarse an instrument to use for a detailed exploration of their distribution in space.

Another method of approach to the problem, which is both powerful and direct, is to make use of the photoelastic effect in transparent crystals. Cubic crystals are particularly suitable, for, since they are not naturally doubly refracting, their birefringence gives a direct picture, subject to certain restrictions, of the magnitude, direction and spatial distribution of the internal stresses.

In the present state of the theory of plastic deformation of crystalline materials the restriction to transparent materials, inherent in this method, is not serious. Since there is no apparent connexion between the plastic properties of a crystalline substance and its transparency or opacity, it is reasonable to assume that results derived from experiments on a transparent material will be applicable to other substances whose mechanism of plastic deformation is essentially similar. In particular, the results should be applicable to metals.

The first observation of artificial birefringence caused by plastic deformation of a crystal seems to have been made in 1815 by Brewster (1818, 1853). The 'lamellar polarisation' that Brewster noticed in rock salt, diamond and certain other cubic crystals was later studied by Biot, but neither experimenter appears to have appreciated that in some of these crystals it was due to plastic deformation (Reusch 1867). More recently, Obreimow & Schubnikoff (1927) have made careful observations of the way the birefringent bands in rock salt appear during the course of plastic deformation, and the nature of the deformation has been further investigated by Brilliantow & Obreimow (1934, 1937) by the use of an ingenious combination of X-ray and optical methods. The photoelastic method does not seem to have been used to explore, in detail, the plastic deformation of any other crystals, nor has it been applied to the internal stresses arising in polycrystalline materials. To approach this problem a substance is needed that is transparent, polycrystalline, ductile, and preferably optically isotropic in the unstressed state. Silver chloride possesses all

the required properties and, in particular, is very ductile at room temperature. Its exceptional qualities have earned it the name of 'transparent metal'.

In this paper we try to show how the characteristic appearance, under the polarizing microscope, of deformed sheets of silver chloride may be explained as the result of internal stresses. There follows a description of experiments to determine the mechanism by which silver chloride deforms plastically, from which we conclude that the process is 'pencil glide'. In part II of this series, with the same general title (Nye 1949), which may be read independently, we give a detailed qualitative explanation of the photoelastic pattern produced between the glide zones in terms of the stresses caused by dislocations.

In addition to the effects of plastic deformation, various optical phenomena make their appearance in the sheets either without stress being applied, or in the elastic stress range. They are interesting as curiosities of the behaviour of crystals in polarized light, but, as most of them have only an indirect connexion with the plastic properties, they are being dealt with in a separate paper.

Short accounts of the experiments have been given in *Nature* (Nye 1948*a, b*) and some of the phenomena were demonstrated at a *Conversazione* held by the Royal Society on 29 May 1947.

2. PREPARATION OF THE SHEETS

The silver chloride used in the experiments was obtained from the Harshaw Chemical Company of Cleveland, Ohio, U.S.A. It is understood that large blocks of the material are formed from the melt under carefully controlled conditions of solidification and cooling, and are then hot rolled. Sheets of thicknesses from 0.3 to 1.0 mm., measuring about 5×5 cm., were supplied. A Leitz Polarizing Microscope, Model CM, was used for the observations. Under ordinary illumination these sheets appeared transparent and colourless, the only marks being scratches on the surface caused by the rolling; but when viewed in white light between crossed nicols, at low magnification, they showed many light areas, for the rolling treatment, as was to be expected, had left them with residual stresses. The sheets were so highly distorted that only rarely could one grain be distinguished from another; the grain boundaries could be made visible, however, by etching for a few seconds in a strong solution of sodium thiosulphate. Most of the experiments to be described were made on strips measuring 30×5 mm. cut from these sheets with a razor blade. The strips were first recrystallized by heating in air for 2 hr. at 400°C (the melting-point of silver chloride is 455°C). After cooling in a furnace (this point will be returned to later in § 9), they were again examined under the microscope between crossed nicols. All the light areas had disappeared and the strips were therefore judged to be free from all but the smallest internal stresses.

The boundaries between the new, recrystallized, grains could be seen on the surfaces of the strips in unpolarized light, when they appeared as fine black lines. The grain size was such that in most places the strips were only one grain thick.

3. THE EFFECT OF PLASTIC DEFORMATION

With sheets of the thicknesses used, the path differences caused by stress birefringence were always less than the wave-length of light. The colour of the restored light was usually white; occasionally, with high loads and thick strips, the first order yellow was reached.

The stress-strain curve of an annealed single crystal of silver chloride, as measured by Stepanow (1934), is similar in shape to that of annealed copper in that the elastic limit (6.9 bars)* is small compared with the stresses that can be withstood by the material after it has been hardened by strain (of the order of 150 bars). Thus, quite a small force in an annealed specimen is sufficient to cause the first trace of plastic deformation; a strip after recrystallization is about as rigid as a piece of cardboard of the same thickness. That a permanent change has taken place can be detected by the fact that, when the external deforming load is taken away, the strip no longer appears completely dark in a polariscope. Figures 1 *a* and *b*, plate 5 show a plastically deformed sheet between crossed nicols. Figure 1 *a* was taken while the external deforming load was still acting; figure 1 *b* was taken after it had been removed.

The light patches that are observed after removal of the load fall broadly into three classes. The majority are irregular in shape and, in general, smaller—though not much smaller—in area than the grains. Some grains, however, are seen to be crossed by straight bands of light. It is possible to see these bands forming while the external stress is being applied. At first they are widely spaced but, on further deformation, new ones form between the old ones and, in addition, grains and parts of grains which before were still in the elastic range are suddenly traversed by a series of these bands. We shall call them ‘birefringent bands’. The third type of residual birefringence is only seen at grain boundaries. The deformed strips contain residual bending moments and, while such a stress system is invisible in the grains themselves, it shows up, for reasons described elsewhere (Nye 1948*a*), at grain boundaries. This bending may consist, in general, of couples acting about two perpendicular axes lying in the plane of the sheet. (If the couples are equal and opposite we have the special case of torsion about an axis in the plane of the sheet.)

Examples of all three types of residual birefringence can be seen in figures 1 *b*, 2 *a*, plate 5 and 3, plate 6. There are several sets of birefringent bands in these photographs, and, in figure 3, more than one set in the same grain. Some of the irregular patches of light have a streaky appearance. Sharp sets of bands are illustrated in figures 4, 5, plate 7 and 6 *b*, plate 8.

If the deformation is continued, fine lines begin to appear on the surfaces of the sheet. In distinction to the birefringent bands and the residual stress patterns referred to above, these lines are visible without the aid of polarized light. They resemble the slip bands seen on metallic surfaces which have been polished and then deformed. Lines of this type are seen in figure 6 *a*, plate 8, which shows the same area as figure 6 *b* but without the analyzer. They show up in reflected as well as in transmitted light and examination under oblique illumination makes it clear that they are fine steps on the surface of the sheet. To estimate their height the ‘shadow-

* 1 bar = 10^6 dynes/cm.².

casting' technique developed for electron microscopy was used. When a beam of gold was made to strike a sheet at a glancing angle of 6° , subsequent measurements under an optical microscope of the widths of the 'shadows' showed that the most prominent steps had a height of about 4000 Å.

It would appear that these steps are due to glide, for, although both the other fundamental mechanisms of plastic deformation, twinning and kinking, can cause a surface to be broken up into facets making angles with each other, they cannot, by their nature, give rise to steps. The plastic behaviour of silver chloride is thus different from that of sodium chloride, which, at room temperature, deforms mainly or entirely by kinking (Brilliantow & Obreimow 1937; Orowan 1942).

If the deformation is by glide it is natural to ask whether the birefringent bands represent the systems of stress set up in the glide packets between successive glide zones. If this is so, the fact that the bands vary considerably in sharpness has a natural explanation; for they should appear sharp only if the glide plane is observed edgewise on. When it is inclined to the direction of the light the different parts of the birefringence pattern in each glide packet will overlap and the details will be obscured. If the angle of inclination is further increased, the stress patterns in different glide packets will overlap and, although the streakiness in the pattern may not be completely obliterated, the details will be quite lost. In agreement with this explanation, it was possible to increase the sharpness of some of the less distinct systems of bands by rotating the specimen under the microscope about an axis parallel to the bands. A Leitz graduated tilting stage was used for the purpose. It was not possible to make the bands in all the grains sharp by rotation, but, since the critical angle for the refraction of light passing from silver chloride into air is only $28^\circ 52'$, this is not surprising. In general, when air is used as the surrounding medium, we could only expect to be able to rotate to sharpness those bands in which the planes are inclined to the normal at angles within the range $-28^\circ 52'$ to $28^\circ 52'$. In practice, owing to mechanical difficulties, this range is considerably reduced.

The connexion between the birefringent bands and the glide lines on the surface must now be discussed. Some grains contain a system of birefringent bands with a series of glide lines running parallel to them. However, bands are frequently seen that are not accompanied by a parallel series of glide lines—this is especially the case in stretched or bent sheets—and, conversely, most glide lines do not run parallel either to a system of bands or to any streakiness in the birefringence pattern. It can be observed, however, that the irregular patches of birefringence which show up most brightly in a circular polariscope are those crossed by the most prominent glide lines. Figures 5, 6 *a* and 6 *b* provide examples of these facts. The set of bands in the upper grain of figure 6 *b* have the same direction as the vertical set of glide lines, but another set of glide lines crosses the same grain in a horizontal direction and does not appear to affect the birefringence pattern to any great extent. On the other hand, no sign whatever could be found of any glide lines following the directions of the sets of bands in figure 5 (although a set of glide lines can be seen crossing them at right angles). In such cases as this, plastic deformation has taken place in a way that leaves no trace on the surfaces; a surface examination of the usual metallographic type would have failed to reveal it.

The explanation of these observations is simple. Birefringent bands without glide lines occur when the glide direction is parallel, and the glide plane is approximately perpendicular, to the surface of the sheet. Glide lines without birefringent bands occur when the glide planes make such large angles with the normal to the sheet that the birefringence pattern is obscured by overlapping. Experiments on rectangular bars (see § 4 (b)), which could be viewed through two surfaces approximately at right angles, confirmed this.

4. EXPERIMENTS TO DETERMINE THE GLIDE ELEMENTS

(a) *Experiments on sheets*

The next step in the investigation was to find the crystallographic indices of the glide plane and glide direction. In the experiments now to be described the crystallographic orientations of six grains showing particularly sharp birefringent band systems were found by taking X-ray Laue photographs. These grains were part of deformed polycrystalline sheets. Two of them are illustrated in figure 5 (strip no. 4) and the other four in figure 4 (strip no. 19). It is unfortunate that the stress system which produced them is not known. The sheets had already been bent between the fingers and then the bands appeared quite suddenly as the strips were being manipulated on a microscope slide. They were possibly the result of a combination of bending and torsion. Band systems of this sharpness rarely occur in strips pulled in tension.

The camera used was a Unicam universal instrument arranged for back-reflexion. The X-ray beam was made sufficiently narrow (~ 0.3 mm.) to fall only on one grain at a time, and this made it possible to take separate back-reflexion Laue photographs of each of the six chosen grains. The X-ray reflexions from some of the grains were sharp, being about 0.5 mm. in width with a crystal to film distance of 25 mm., but often they were drawn out to lengths of 1 to 2 mm. Occasionally the spots were noticeably double, with separations of from 0.5 to 1 mm.; this corresponds to rotations of about 0.5 to 1° . The indices of the spots were found by the method described by Greninger (1935), and a stereographic projection was plotted for each grain showing the positions of the poles of the crystallographic planes in relation to the faces and edges of the strip in which it was situated. Each grain in turn was then rotated under the microscope by means of the tilting stage until the bands in it appeared to be most clearly resolved. By noting the angle of tilt and allowing for refraction at the surface the inclination of the birefringent lamellae was calculated. In this way the pole, P , of the glide plane was plotted on the stereographic projection. Table 1 gives the various angular relationships for each set of bands, derived with the aid of a stereographic net of radius 15 cm. Each line of the table refers to one set of birefringent bands, specified in column 3. In column 4 is entered the crystallographic plane of low indices having its pole nearest to P . The agreement between glide plane and crystal plane can best be judged from the angles α and β in columns 5 and 6. The figures given for α show that the glide planes are in each case (except for the doubtful set C in grain 1 of strip 19) close to either a $\{111\}$ plane or a $\{112\}$ plane. If the planes were really coincident with these crystallographic planes, however, the

accuracy of the experiments would lead to angles of 1 or 2° at the most for α . The lack of close agreement between the angles γ and δ also shows that the glide plane is not exactly crystallographic.

TABLE 1. THE CRYSTALLOGRAPHIC ORIENTATIONS OF THE GLIDE PLANES GIVING BIREFRINGENT BANDS IN PLASTICALLY DEFORMED SHEETS OF SILVER CHLORIDE

1	2	3	4	5	6	7	8	9	10	11
strip no.	grain no.	set of birefringent bands	nearest crystal plane	angles in degrees						surface lines in same direction as bands
				α	β	γ	δ	ζ	ϵ	
4	A	—	(111)	8.1	0.0	3.3	0.2	0.5	6.5	no
4	B	—	(111)	3.0	0.3	3.1	0.1	0.0	5.8	no
19	1	A (short bands)	(111)	2.2	1.0	4.2	6.2	0.5	42.0	yes
19	1	B (long bands)	(112)	2.3	2.0	2.2	4.1	1.0	28.0	yes
19	1	C*	(100)	4	0.7	9.1	5	0.5	11.0	yes
19	2	A	(111)	2.0	2.0	2.0	2.1	2.0	4.2	no
19	2	B†	(112)	4.5†	4.3	14	14.8†	5.0†	16.2	faint trace
19	3	—	(111)	2.7	0.0	7.5	4.8	0.5	4.9	no
19	4	—	(111)	4.5	1.3	3.5	1	1.2	3.4	no

α = angle between P and the pole of the plane given in column 4.

β = angle between bands and the trace which the plane given in column 4 makes on the surface.

γ = angle between normal to the surface and the plane given in column 4.

δ = angle between normal to the surface and the glide plane.

ζ = angle between glide plane and the nearest $\langle 110 \rangle$ direction.

ϵ = angle between this $\langle 110 \rangle$ direction and the surface.

* This set faint and barely measurable.

† Measurements of δ , and therefore of α and ζ , were rough for this set.

An interesting light is thrown on this fact by the values of the angle ζ given in column 9. ζ is the measured angle between the glide plane and the $\langle 110 \rangle$ direction lying closest to it. It will be noticed that the unusually high value of 5.0° occurs for a set of bands on which only rough measurements could be made, while the rest of the figures in this column are less than, or equal to, the experimental error. We

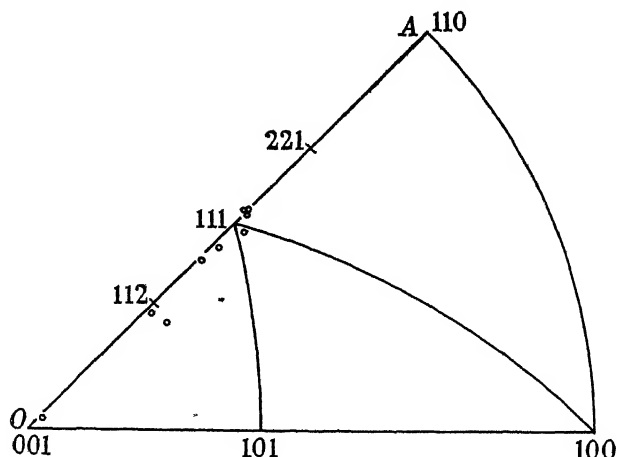


FIGURE 8. Stereographic projection showing the poles of the glide plane in the grains of silver chloride sheets.



FIGURE 1a.



FIGURE 1b.

FIGURE 1a. Silver chloride sheet deformed plastically by tension applied in horizontal direction. Crossed nicols. Vibration directions parallel to cross-wires. (Magn. $\times 23$.)

FIGURE 1b. The same but with external tensile load removed. (Magn. $\times 23$.)



FIGURE 2a.

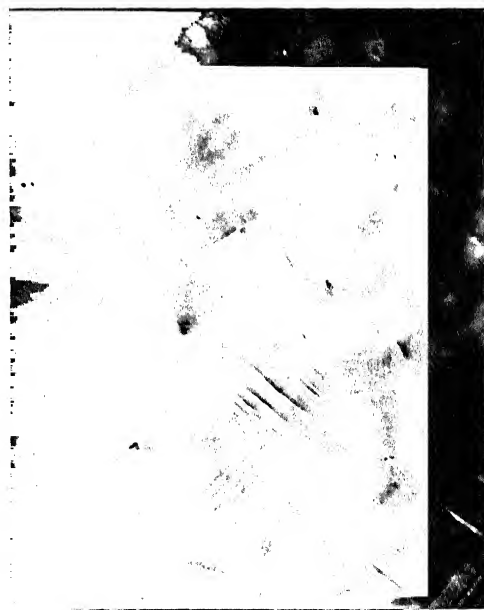


FIGURE 2b.

FIGURE 2a. Silver chloride sheet after plastic deformation. Crossed nicols. Vibration directions at 45° to cross-wires. (Magn. $\times 19$.)

FIGURE 2b. The same after heating the sheet to 355°C and allowing to cool in the furnace. (Magn. $\times 19$.)

(Facing p. 196)

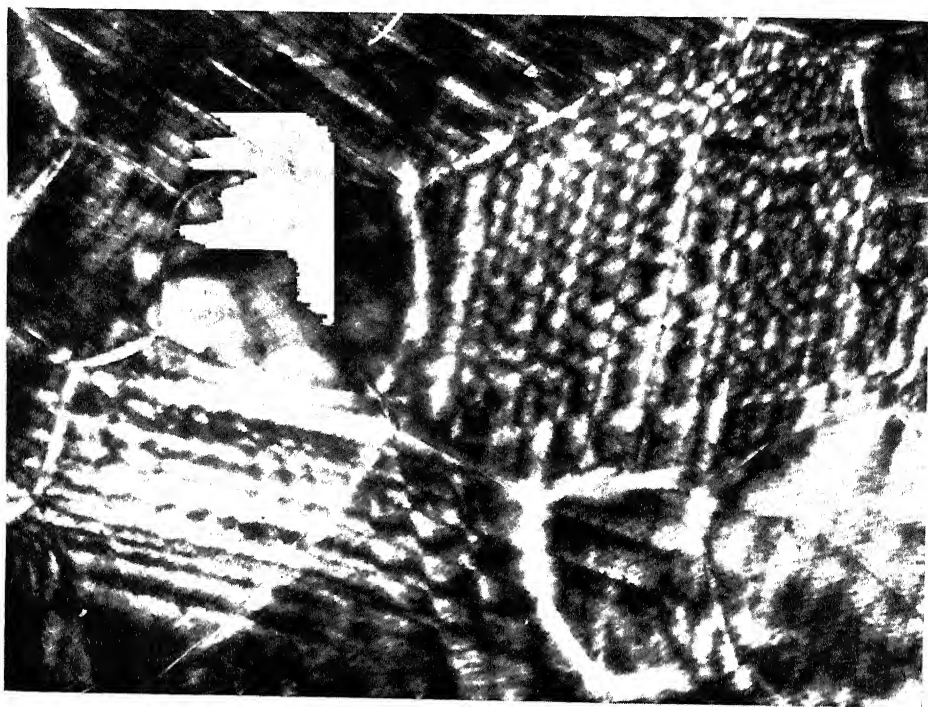


FIGURE 3. Intersecting systems of birefringent bands. Crossed nicols. (Magn. $\times 63$.)

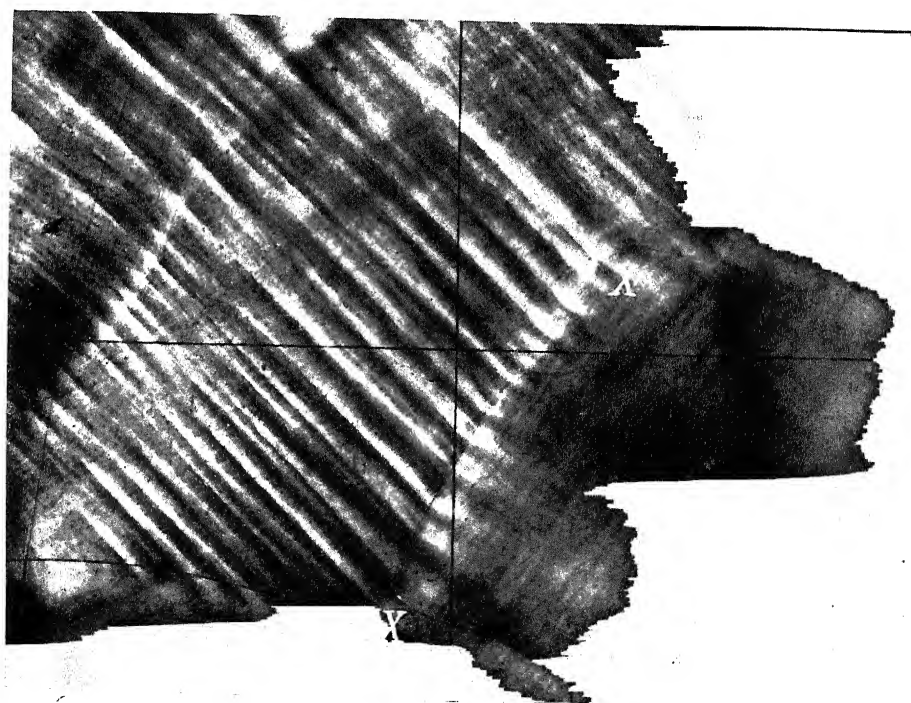


FIGURE 7. Birefringent bands. Crossed nicols. (Magn. $\times 94$.)



FIGURE 4. Strip no. 19. Measurements on grains 1, 2, 3 and 4 are given in table 1. (Magn. $\times 25$.)

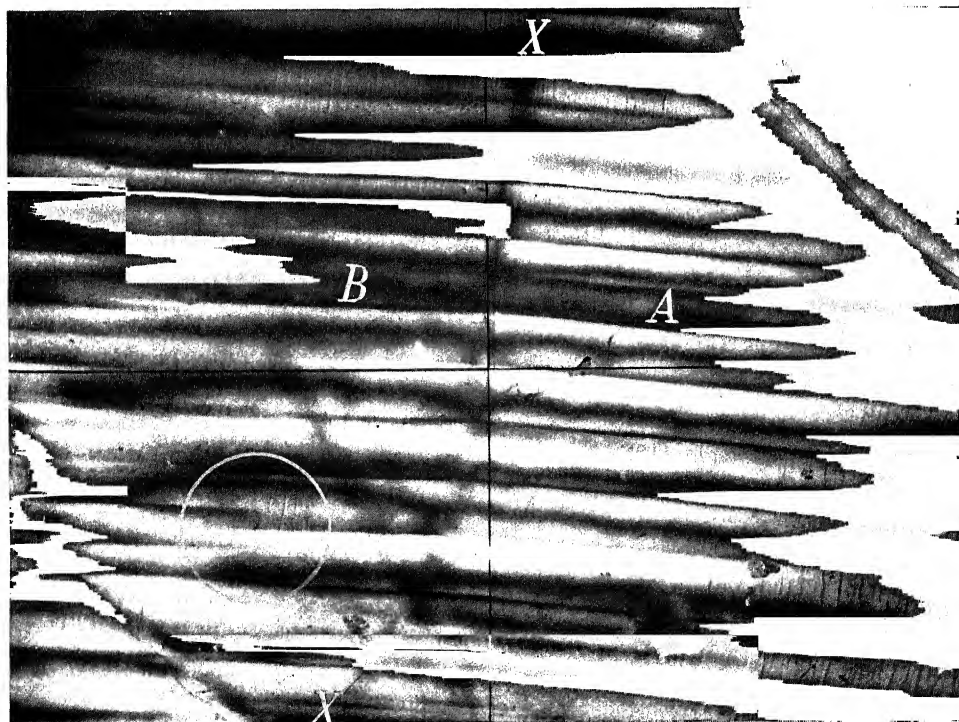


FIGURE 5. Strip no. 4. Measurements on the bands in grains A and B are given in table 1. (Magn. $\times 147$.)

FIGURES 4, 5. Birefringent bands in sheets of silver chloride seen between crossed nicols. Vibration direction of nicols parallel to cross-wires.

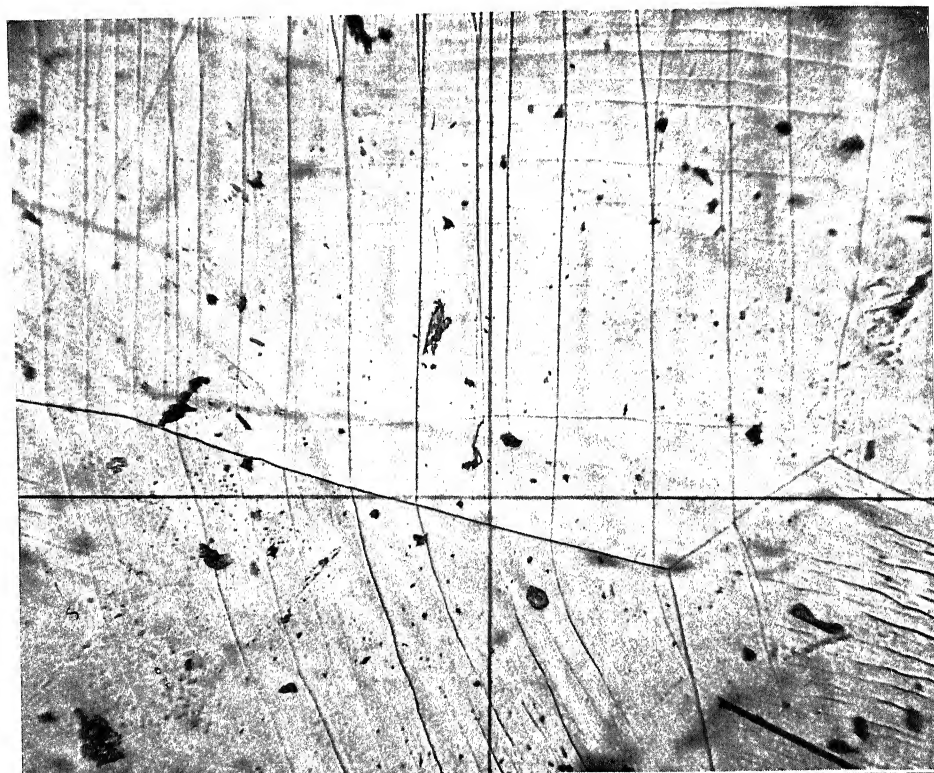
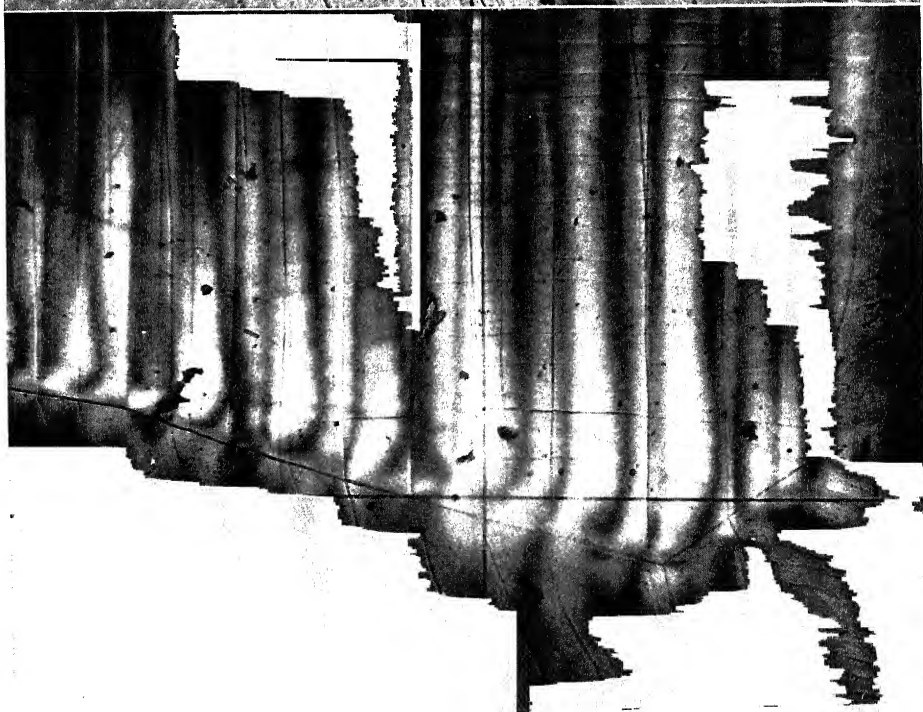
*a**b*

FIGURE 6*a*. Showing glide lines and grain boundaries on the surface of a silver chloride sheet. Portions of four grains are visible. Polarizer in position but no analyzer. (Magn. $\times 97$.)

FIGURE 6*b*. The same as figure 6*a* but seen in a circular polariscope. (Magn. $\times 97$.)

conclude that, although the glide planes do not coincide with crystallographic planes, they always contain a $\langle 110 \rangle$ direction.

This is shown graphically by the stereographic projection in figure 8. The plotted points are the poles of the glide planes. The $\langle 110 \rangle$ direction lying closest to the glide plane has in each case been assigned the indices $[\bar{1}10]$ and all the points have been arranged to fall within one octant of the primitive circle. Such a procedure amounts only to a suitable choice of x , y and z axes in each grain. (When a grain contains more than one system of glide bands the x , y and z axes have to be relabelled to deal with each set.) It will be seen that the points do not cluster in one position but are distributed near the great circle (straight line) OA , which is the locus of the poles of planes lying in the $[\bar{1}10]$ zone.

The results shown in the last two columns of table 1 will be referred to later.

(b) *Experiments on rectangular bars*

The experiments of which details are given in table 1 had necessarily to be confined to grains containing band systems that could be made sharp by tilting the sheet, for only in these cases was it possible to determine the true orientation of the glide plane. Experiments have also been made on specimens in which the true orientation of the glide planes could be found by measuring the direction of their traces on two non-parallel surfaces, regardless of whether any birefringent bands could be made visible or not. These specimens were in the form of rectangular bars of nearly square cross-section (measuring approximately $25 \times 1 \times 1$ mm.) cut from a rolled sheet. After the four main surfaces had been polished by a microtome method, the bars were recrystallized by heating for 2 hr. at 400°C , cooled slowly in the furnace, and then the scratches left by the microtome were removed by dipping the bars for 15 sec. in a strong solution of sodium thiosulphate. This treatment revealed the boundaries of the recrystallized grains, and considerable portions of the specimens were seen, under the microscope, to consist of single chains of large grains. That is to say, each bar was equivalent to a succession of single crystals joined up end to end. A further series of back-reflexion Laue photographs was taken to find the crystallographic orientation of some of the glide planes produced by extension of these bars. Since the direction of the tension was known, such experiments also gave the possibility of studying the relation between the glide plane and the stress system. The results are shown in tables 2 and 3.

Six sets of active glide planes were examined. The four grains in which they lay were all from two to three times as long as they were broad, the lengths being measured parallel to the axes of the bars. The tension in the central parts of these grains where the measurements were made would therefore have been reasonably uniform. The pole of each set of glide planes was found, with the aid of a stereographic net, by measuring the direction of the traces on at least two surfaces (non-parallel). In some cases the directions of the traces on all four surfaces were measured. The X-ray photographs were used, as before, to plot the positions of the poles of the crystallographic planes on the same stereographic projection. For the first four sets of glide planes given in table 2 the nearest crystallographic plane of low indices was of the form $\{111\}$; values of the angle, α , between this plane and the observed glide plane

TABLE 2. THE CRYSTALLOGRAPHIC ORIENTATION OF GLIDE BANDS IN TENSILE BARS OF APPROXIMATELY SQUARE CROSS-SECTION

1 bar no.	2 grain no.	3 set of glide bands	4 nearest crystal plane	angles in degrees		
				α	ζ	η
3	1	A	(111)	6.0	0.0	3.0
3	2	A	(111)	4.0	0.5	2.0
3	2	B	(111)	6.3	1.5	1.8
4	1	A	(111)	4.0	0.2	8.9
4	2	A	{(111) {(221)	11.4 4.6	0.0	0.7
4	2	B	{(111) {(221)	11.8 4.7	2.3	6.7

α = angle between glide plane and the nearest crystal plane (column 4).

ζ = angle between glide plane and the $\langle 110 \rangle$ direction lying closest to it.

η = angle between glide plane and the plane of maximum resolved shear stress containing the glide direction.

are given in column 5. The last two sets were nearer to $\{221\}$ than $\{111\}$ planes, but the angles with both planes are given in the table for comparison. The experimental error in the determination was due mainly to variation in the directions of some of the slip lines across the grains and to the distortion of the faces and edges of the grains; the latter caused uncertainty in the setting of the specimens in the X-ray camera and on the rotating and tilting stage of the microscope. The errors from these sources would be expected to be about $\pm 1.5^\circ$. The angles given in column 5, therefore, make it certain that the glide planes, although near to crystallographic planes, are again not actually coincident with them. Nevertheless, in agreement with the experiments on the sheets, the values of ζ in column 6 are all much smaller than those of α and it

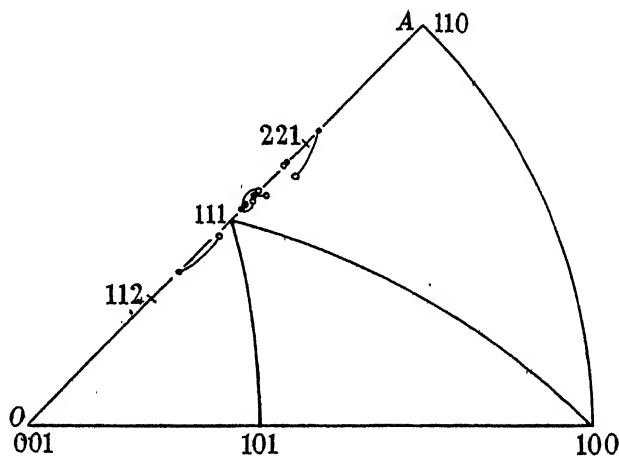


FIGURE 9. Stereographic projection showing the orientation of the glide planes in the rectangular bars. \circ Pole, P , of glide plane. \bullet Pole, P' , of plane containing the glide direction and subject to the maximum resolved shear stress.

seems permissible to conclude that the glide plane, although not accurately $\{111\}$ or $\{221\}$, always contains a $\langle 110 \rangle$ direction.

Figure 9 is plotted on the same basis as figure 8. The open circles, which show the poles of the glide planes, are again distributed close to OA .

(c) Pencil glide

These results suggest a comparison between the plastic behaviour of silver chloride and that of α -iron. Taylor & Elam (1926) found that in α -iron at room temperature, although the direction of glide was crystallographically determined, the glide plane was not; it depended on the state of stress. This peculiar behaviour of iron has been studied by other investigators, including Barrett, Ansel & Mehl (1937), and appears to be connected with the fact that the $[111]$ direction is common to three different glide planes, not crystallographically equivalent, but each having the same critical shear stress at room temperature. Glide takes place in the $\langle 111 \rangle$ direction with a movement like that obtained by shearing a bundle of pencils in the direction of their length.

In view of this the natural course was to find whether the $\langle 110 \rangle$ direction which lay in the glide plane of silver chloride was in fact the glide direction. We find indications of this in the experiments on the sheets. Column 10 of table 1 gives the angle, ϵ , which this $\langle 110 \rangle$ direction makes with the surface of the sheet. If it is the glide direction there should be no surface steps when it lies in the plane of the sheet and, other things being equal, the steps would be expected to increase in height as it makes increasingly large angles with the surface. It can be seen, by comparing columns 10 and 11, that whenever ϵ is small there are no surface lines running in the same direction as the birefringent bands and, conversely, whenever ϵ is large surface lines are observed. The hypothesis that $\langle 110 \rangle$ is the glide direction is thus strengthened.

It is worth remarking that, since the $\langle 110 \rangle$ direction in a crystal of silver chloride is parallel to rows of like ions, it would be surprising if it were not the glide direction.

A satisfactory theory of pencil glide has still to be found and this is not the place for speculation on the subject. We will content ourselves by citing the work of Elam (1936) on β -brass and that of Andrade and his collaborators on slip in the body-centred cubic metals (Andrade 1938; Andrade & Chow 1940).

5. THE RELATION BETWEEN THE GLIDE ELEMENTS AND THE APPLIED STRESS

We now come to the relation between the direction of the tension and the glide elements in the rectangular bars. We first investigate the connexion between the axis of tension and the $\langle 110 \rangle$ direction lying in the glide plane. (A plane of maximum shear stress makes an angle of 45° with the tension and the direction of the shear stress in such a plane also makes 45° with the tension axis.) Table 3 gives the values of $|\theta - 45^\circ|$, where θ is the angle between a $\langle 110 \rangle$ direction and the axis of tension, for all the crystallographically equivalent $\langle 110 \rangle$ directions in the four grains whose orientations were known from the X-ray photographs. The values referring to the $\langle 110 \rangle$ directions that lie in the observed glide planes, and which we have tentatively

assumed to be the operative glide directions, are shown in heavy type; in each case they are seen to be the lowest values in their column. (For the purposes of table 3 an arbitrary choice of x , y and z axes was made in each grain. The axes are therefore not those used in plotting the poles in figure 9, where the axes were always arranged so as to make $[\bar{1}10]$ the glide direction.)

TABLE 3. SHOWING HOW THE SIX $\langle 110 \rangle$ DIRECTIONS ARE RELATED TO THE DIRECTION OF THE TENSION

direction	values of $ \theta - 45^\circ $ (in degrees)			
	bar no. 3		bar no. 4	
	grain 1	grain 2	grain 1	grain 2
$[110]$	13.7	23.8	29.0	10.4 <i>B</i>
$[\bar{1}\bar{1}0]$	6.0 <i>A</i>	23.4	22.7	6.5 <i>A</i>
$[101]$	23.5	1.8 <i>A</i>	6.8 <i>A</i>	15.2
$[10\bar{1}]$	38.9	7.8 <i>B</i>	20.8	13.4
$[011]$	20.6	31.1	37.2	28.5
$[01\bar{1}]$	17.8	26.5	14.3	41.2

Angles corresponding to active glide directions are shown in heavy type and labelled with the letter corresponding to the glide band system.

We are thus led to the view that glide takes place in those $\langle 110 \rangle$ directions which lie most nearly at 45° to the axis of tension.

The glide plane must, of course, contain the glide direction, and in Taylor & Elam's work on iron it was found to be close to that one, out of the zone of planes containing the glide direction, on which the maximum shear stress, resolved in the glide direction, acted. In a crystal under uniaxial tensile stress σ the shear stress, τ , on a glide plane resolved in the glide direction is given by

$$\tau = \sigma \cos \phi \cos \lambda,$$

where ϕ is the angle between the normal to the glide plane and the axis of tension, and λ is the angle between the glide direction and the axis of tension.

If λ and σ are fixed, τ is a maximum when ϕ is a minimum. But ϕ is least when the normal to the glide plane lies in the same plane as the glide direction and the axis of tension. This determines the position of the glide plane in an ideal case in which the above criterion is exactly satisfied.

We may test how closely the condition is satisfied by constructing the pole, P' , of this ideal plane. The positions of P' for each of the six glide systems are shown by the filled-in circles in figure 9. The results of measuring the angle, η , between P' and P , the pole of the observed glide plane, are tabulated in column 7 of table 2. It will be seen that the correspondence between P and P' is not exact, but nevertheless quite close, the maximum divergence being 8.9° . It is apparent from figure 9 that the observed glide plane tends to deviate from the position of maximum resolved shear stress towards a position lying between the (221) and (111) planes. It is therefore likely that the critical resolved shear stress for glide is not quite the same for all the planes in the $[\bar{1}10]$ zone, and that it is a minimum for a plane somewhere between (221) and (111). This may explain why the points

in figure 9 are only found near the central portion of the line OA ; for the grains with orientations that would have given glide planes at other points of OA were 'harder' than those that gave poles in the central region; and, since only grains with well-developed glide lines were chosen for measurement, these hard grains, which may not have slipped at all, were missed from the sample.

The only other determination of the glide elements in silver chloride known to me is that of Stepanow (1934, 1935). When single crystals were extended by simple tension he found glide in a $\langle 110 \rangle$ direction on planes within 5° of $\{110\}$. This result is in agreement with the conclusions reached here, for in Stepanow's experiments the crystals happened to be prepared with a $\{100\}$ plane perpendicular to the axis of tension and, with this particular orientation, it is easy to see that the possible glide plane on which the maximum resolved shear stress acts is crystallographic— $\{110\}$ in fact.

6. THE CORRUGATED NATURE OF THE GLIDE SURFACES

It has been mentioned that in flat strips of silver chloride either bent or extended by tension birefringent band systems frequently occurred unaccompanied by any glide lines. The reason for this is now clear. The planes of maximum shear stress envelop a cone of semi-vertical angle 45° whose axis is the direction of the tension (or compression). Out of this group of planes, those which lie perpendicular to the plane of the sheet will tend to glide in a direction parallel to the surface of the sheet.

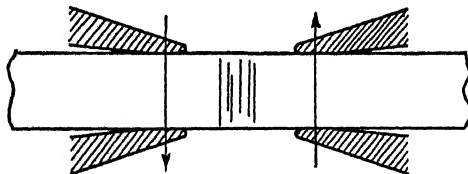


FIGURE 10. Deformation of silver chloride bar by simple shear.

There will thus be a tendency for sharp band systems without surface steps to form at 45° to the axis of tension. When glide lines do occur with such band systems they are invariably straight. This is in marked contrast with the wavy glide lines seen in grains in which the slip direction makes a considerable angle with the intersection of glide plane and surface. These observations indicate that the glide surface is not quite plane, but is corrugated in such a way that sections parallel to the glide direction are straight lines, while sections perpendicular to this are lines of a fixed general direction, but wavy in detail. Similar irregularities in the glide surfaces have been observed in other materials, notably in α -iron (Taylor & Elam 1926), although the phenomenon in this metal may depend on the impurities (Andrade 1938), and under certain conditions in mercury (Greenland 1937).

Some other experiments confirmed this corrugated nature of the glide surfaces. When rectangular bars of the type already described were deformed by simple shear by holding them in tweezers in the way shown in figure 10, they sometimes gave glide planes perpendicular to the axis of the bar. This only happened in favourably oriented crystals, for one would not expect that a $\langle 110 \rangle$ direction would always

be found in the direction of shear. However, the experiment could be performed on many different crystals in the bar, and when nine different bars were tested in this way a number of such glide systems were produced. In all cases the glide lines in the direction of slip were quite straight, while those on the face perpendicular to the direction of slip were wavy and in many cases divided into several branches.

It may be noted at this point that, on the theory of pencil glide presented by Taylor & Elam and adopted here, the observation of glide surfaces that approximate to planes is a consequence of a uniform plastic strain in the crystals. (The small scale non-uniformity introduced by the fact that glide does not take place on every atomic plane, but is concentrated in zones, is a much smaller effect than those we are now considering.) If glide took place by equal amounts on a series of equally spaced cylindrical surfaces of more complicated type than planes, the resulting large-scale distortion of the crystal would not be a uniform simple shear. Conversely, a non-uniform distribution of stress would be expected to produce curved glide lines on the surface of the crystal. Such curved glide lines could always be seen in a silver chloride sheet near places where the stress had been inhomogeneous, near an indentation, for instance, or in the distorted region left by the razor blade used for cutting the sheet. Each glide line corresponds, at all points of its length, to one glide direction, and it terminates at the points in the crystal where this direction ceases to be the one corresponding to the most favourable plane for slip.

7. EFFECT OF GRAIN BOUNDARIES ON SLIP PROPAGATION

On the hypothesis of pencil glide, the progress of plastic deformation in a polycrystalline specimen of silver chloride (or α -iron) might be expected to occur as follows. At first all grains are within their elastic limits. When the stress is increased, the critical resolved shearing stress for glide is reached first on a particular set of planes in one particular grain, which we denote by A . This grain will slip and, since its surface is now stepped, it will exert non-uniform stresses on its neighbours B , C , D , etc. If the applied stress is further increased, the part of B , for instance, which is situated near the surface of contact with A will either remain purely elastically strained, or will slip plastically in such a way as to conform to the stepped surface of A . In a material which was capable of glide only on crystallographic planes this could not be achieved in a simple way for the well-known reason that no potential glide plane of B would intersect an operative glide plane of A in a line that lay in the grain boundary. In crystals capable of pencil glide, however, conformity at the boundaries can be achieved without difficulty. The glide surface in the grain B , for instance, would be determined by the following construction. If XY is the line of intersection of one of the operative glide planes in A with the boundary between A and B , draw through each point of XY lines whose directions are the possible glide directions in B , six in number if the glide direction is $\langle 110 \rangle$ (and four if it is $\langle 111 \rangle$). These lines will be the generators of six (or four) surfaces on which glide is geometrically possible. The surfaces will be planes if the grain boundary is itself a plane and glide in A has taken place on a plane, but, in general, they will be cylindrical surfaces of complicated form. In the immediate neighbourhood of the

boundary the resolved shear stress acting on each of these glide systems will be largely due to the alternating stresses caused by contact with the stepped surface of *A*. That one will operate on which the resolved shearing stress is greatest, but it may, of course, be necessary to raise the external stress before this happens. In the centre of *B* the stress system will be disturbed from the configuration it had before grain *A* slipped not so much by these alternating stresses as by the macroscopic change of shape which grain *A* has suffered during the slip process. In general, then, a different slip system will tend to operate in the centre of the grain from that needed near its surface. The same considerations apply to *A*'s other neighbours *C*, *D*, etc., and similar processes will occur throughout the whole specimen.

These views, while partly conjectural, find support in the present experiments. For instance, the grain immediately below the intersection of the cross-wires in figure 2*a*, plate 5 contains a sharp set of birefringent bands and at the same time its neighbour on the left is stressed non-uniformly down its right-hand edge by the translations that have taken place parallel to the bands. The same effect is seen in figure 7, plate 6 at the boundary marked *XX*. (For these photographs, the external stress was removed.) Many other instances of this effect have been seen.

Again, the frequent occurrence of glide lines which do not stop at grain boundaries but continue across them without a break is a natural consequence of the property of pencil glide. It is difficult to conceive such an event, except as a coincidence, in crystals not capable of pencil glide. The effect is well seen in figure 6*a*, plate 8. Another instance of a similar event is provided by figure 5, plate 7. Here there are two grains *A* and *B* separated by the boundary *XX*. The birefringent bands in each grain are in step as they cross *XX*, but it can be seen that they turn through a sharp angle. Referring to the values of δ in lines 1 and 2 of table 1, we see that both sets of birefringent lamellae are very nearly perpendicular to the surface. (This is also approximately true of the boundary.) This again would be inexplicable, except as a coincidence, with the usual glide mechanism.

It may also be noticed that grain *B* in this photograph contains several glide zones that end in the centre of the grain. One example is marked with a circle.

8. TYPES OF RESIDUAL STRESS WITHIN GRAINS

The main types of residual stress seen in the silver chloride sheets may be briefly listed.

(1) A system of stresses set up between the glide zones of each grain and alternating with a period equal to the spacing of the glide zones. A detailed analysis of these is given in part II of this series (Nye 1949). They are produced in both single crystal and polycrystalline specimens and can reasonably be attributed to the presence of dislocations trapped in the glide zones.

(2) The alternating stresses produced near the grain boundaries at each end of a system of glide zones in a slipped grain.

(3) 'Heyn stresses' produced by non-uniformity of plastic deformation from grain to grain (see, for example, Seitz 1943). (This category includes the residual bending moments referred to in § 3.)

9. THE ANNEALING PROCESS

A series of experiments has been performed with the object of studying the progress of annealing in a cold worked silver chloride sheet. The thermal expansion of cubic crystals is isotropic and, provided the temperature of the sheet were always uniform, no stresses due to the different expansions of the grains, such as are found in polycrystalline specimens of non-cubic materials, would be expected.

The sheet shown in figure 2*a* was heated to 355° C and allowed to cool in the furnace. Under these conditions recrystallization did not occur and the sheet appeared as shown in figure 2*b*. There was no alteration in the position of the light places, but only a reduction in their intensity. This shows that annealing does not bring about any large-scale redistribution of stress, but only a general reduction of its magnitude. After a further period of heating of $\frac{3}{4}$ hr. at 415° C the pattern had almost disappeared. It was essential to cool slowly from the high temperature if a strip was to be obtained free from internal stresses. The thermal gradients set up by removing a specimen directly from the furnace at 400° C to the air of the room caused internal stresses and plastic deformation. It is true that silver chloride is a bad conductor of heat, but the photoelastic method is more sensitive than a surface examination in revealing plastic deformation. This experiment illustrates the care that must be taken in the annealing even of cubic materials if a specimen free from internal stresses is to be obtained at room temperature.

The use of silver chloride for the investigations described in this part and part II was suggested by Dr E. Orowan. I should like to thank him warmly for allowing me to develop his idea and for his help and counsel during the work.

The experiments were financed by the British Iron and Steel Research Association.

REFERENCES

- Andrade, E. N. da C. 1938 *Proc. Roy. Soc. A*, **168**, 310.
 Andrade, E. N. da C. & Chow, Y. S. 1940 *Proc. Roy. Soc. A*, **175**, 290.
 Barrett, C. S., Ansel, G. & Mehl, R. F. 1937 *Trans. Amer. Soc. Metals*, **25**, 702.
 Brewster, D. 1818 *Trans. Roy. Soc. Edinb.* **8**, 157.
 Brewster, Sir David 1853 *A treatise on optics*, pp. 280, 281. London: Longman, Brown, Green and Longmans.
 Brilliantow, N. A. & Obreimow, I. W. 1934 *Phys. Z. Sowjet.* **6**, 587 (in English).
 Brilliantow, N. A. & Obreimow, I. W. 1937 *Phys. Z. Sowjet.* **12**, 7 (in English).
 Elam, C. F. 1936 *Proc. Roy. Soc. A*, **153**, 273.
 Greenland, K. M. 1937 *Proc. Roy. Soc. A*, **163**, 28.
 Greninger, A. B. 1935 *Trans. Amer. Inst. Min. (Metall.) Engrs*, **117**, 61.
 Nye, J. F. 1948*a* *Nature*, **161**, 367.
 Nye, J. F. 1948*b* *Nature*, **162**, 299.
 Nye, J. F. 1949 In course of preparation.
 Obreimow, I. W. & Schubnikoff, L. W. 1927 *Z. Phys.* **41**, 907.
 Orowan, E. 1942 *Nature*, **149**, 643.
 Reusch, E. 1867 *Ann. Phys., Lpz.*, (v), **132**, 441; *Proc. Roy. Soc. Edinb.* **6**, 134.
 Seitz, F. 1943 *The physics of metals*, p. 147. New York: McGraw-Hill.
 Stepanow, A. W. 1934 *Phys. Z. Sowjet.* **6**, 312 (in English).
 Stepanow, A. W. 1935 *Phys. Z. Sowjet.* **8**, 25 (in German).
 Taylor, G. I. & Elam, C. F. 1926 *Proc. Roy. Soc. A*, **112**, 337.

One-dimensional dislocations.

I. Static theory

By F. C. FRANK AND J. H. VAN DER MERWE

H. H. Wills Physical Laboratory, University of Bristol

(Communicated by N. F. Mott, F.R.S.—Received 22 December 1948—

Revised 25 March 1949—Read 19 May 1949)

The theory of a one-dimensional dislocation model is developed. Besides acting as a pointer to developments of general dislocation theory, it has a variety of direct physical applications, particularly to monolayers on a crystalline substrate and to conditions in the edge row of a terrace of molecules in a growing crystal. Allowance is made in the theory for a difference in natural lattice-spacing between the surface layer or row and the substrate. The form and energy of single dislocations and of regular sequences of dislocations are calculated. Critical conditions for spontaneous generation (or escape) of dislocations are determined, and likewise the activation energies for such processes below the critical limits. Various physical applications of the model are discussed, and the physical parameters are evaluated with the aid of the Lennard-Jones force law for the above-mentioned principal applications.

1. INTRODUCTION

The theory of dislocations is moderately complex even with extensive approximations and unreal restrictions such as that to a simple cubic lattice, and there remain many problems, e.g. of dislocation dynamics, which have only been treated qualitatively even with these simplifications. It has therefore seemed worth while to us to make a more extensive survey of the properties of dislocations in a still more drastically simplified model, the simplest which can display dislocations at all. There are several different mechanical systems, all leading to the same equations. We choose one of them to set up the fundamental equation. It consists of a number of identical balls connected by identical springs and arranged in a straight line to which their motion is constrained, the balls being at the same time acted on by a force which varies periodically along the straight line. In the first instance, we take the periodic field to be sinusoidal. This is realized to a close approximation if the chain of balls and springs rests in a long horizontal frictionless trough with vertical side walls and a sinusoidal corrugation of low amplitude on the bottom of the trough, so that the periodic field is provided by gravity. We may hereafter refer indiscriminately to the balls as 'atoms' and to the source of the periodic field as the 'substrate', though this does not express the only application of the equations to the physics of solids. Certain aspects of this or equivalent systems have been treated by Dehlinger (1939), Frenkel & Kontorowa (1938) and Lennard-Jones (1940), but for brevity we derive the results *ab initio*.

2. THE DIFFERENCE EQUATION

If a is the 'wave-length' of the potential due to the substrate and $\frac{1}{2}W$ its amplitude, b is the spacing between atoms of the chain when their connecting springs are unstrained, μ is the force constant of these springs, and x_n is the displacement of the

n th atom from the n th trough of the substrate potential, each numbered in sequence from an arbitrary commencement, then the potential energy corresponding to N atoms of the chain is

$$V_N = \frac{1}{2}\mu \sum_{n=0}^{N-1} (x_{n+1} - x_n + a - b)^2 + \frac{1}{2}W \sum_{n=0}^{N-1} (1 - \cos 2\pi x_n/a).$$

This may be written

$$V_N = Wl_0^2 \sum_0^{N-1} (\zeta_{n+1} - \zeta_n - 1/P_0)^2 + \frac{1}{2}W \sum_0^{N-1} (1 - \cos 2\pi \zeta_n), \quad (1)$$

if we introduce the abbreviations

$$\zeta_n = x_n/a, \quad P_0 = a/(b-a), \quad l_0 = (\mu a^2/2W)^{\frac{1}{2}}.$$

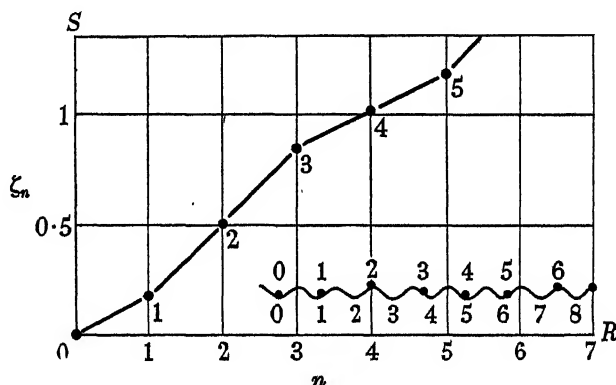


FIGURE 1

ζ is here the displacement measured in units of the substrate spacing; P_0 may be called the vernier period of misfit; the significance of l_0 will become evident later. The condition for equilibrium of the n th atom is $\partial V/\partial \zeta_n = 0$, whence by differentiation of (1)

$$\Delta_n^2 \zeta = (\pi/2l_0^2) \sin 2\pi \zeta_n. \quad (2)$$

Here we employ central difference notation

$$\Delta_n^2 \zeta = \zeta_{n-1} - 2\zeta_n + \zeta_{n+1}.$$

Solutions of equation (2) can be constructed numerically from specified displacements of any two successive atoms. Some simple symmetrical solutions can be found by elementary algebra, e.g. that of figure 1 corresponding to

$$1/l_0^2 = (1 - 4\zeta_1)/\pi \sin 2\pi \zeta_1 \quad (0 < \zeta_1 \leq \frac{1}{4}).$$

3. THE DIFFERENTIAL EQUATION AND ITS SOLUTIONS

When ζ_{n+1} and ζ_{n-1} are expanded in Taylor series about the point n one obtains

$$\Delta_n^2 \zeta = \zeta_{n+1} - 2\zeta_n + \zeta_{n-1} = \frac{d^2 \zeta}{dn^2} + \frac{2}{4!} \frac{d^4 \zeta}{dn^4} + \frac{2}{6!} \frac{d^6 \zeta}{dn^6} + \dots$$

When l_0 is large, i.e. the springs are relatively strong, $\Delta_n^2 \zeta = (\pi/2l_0^2) \sin 2\pi\zeta_n$ will be small, and one is justified to a first approximation in neglecting differential coefficients of higher orders. Hereafter we shall use, for the condition of equilibrium, the differential equation

$$d^2\zeta/dn^2 = (\pi/2l_0^2) \sin 2\pi\zeta, \quad (3)$$

and keep in mind that the greater l_0 the better the approximation.

Integration of (3) gives

$$\begin{aligned} (d\zeta/dn)^2 &= \epsilon^2 + (1 - \cos 2\pi\zeta)/(2l_0^2) \\ &= \frac{1 + \epsilon^2 l_0^2}{l_0^2} \left(1 - \frac{\cos^2 \pi\zeta}{1 + \epsilon^2 l_0^2} \right), \end{aligned} \quad (4)$$

ϵ^2 being an integration constant. $d\zeta/dn$ is the amount by which the distance between successive atoms, measured in units of a , exceeds unity. ϵ is evidently the value of this excess where $\zeta = 0$, i.e. where successive atoms are in or close to troughs of the substrate potential.

Let us write

$$k^2 = 1/(1 + \epsilon^2 l_0^2) \quad \text{and} \quad \phi = \pi(\zeta - \tfrac{1}{2}),$$

and choose the zero for n where $\zeta = \frac{1}{2}$, i.e. $\phi = 0$, then (4) becomes

$$d\phi/dn = \pm (\pi/l_0 k) (1 - k^2 \sin^2 \phi)^{\frac{1}{2}}, \quad (5)$$

where the +ve sign refers to expansions of the chain of atoms with respect to the substrate, producing what we shall call negative dislocations. The negative sign corresponds to compression relative to the substrate, and positive dislocations. If we introduce the convention that $l_0 > 0$ for negative dislocations and $l_0 < 0$ for positive ones, we may discard the \pm sign in (5) without loss of generality. It is convenient to consider negative dislocations, which we shall do hereafter, and keep in mind that the results obtained are also valid for positive dislocations.

(i) *Dislocations far apart* ($\epsilon = 0$; $k = 1$)

If $\epsilon = 0$, then $k = 1$. Then from (4)

$$d\zeta/dn = (1/l_0) \sin \pi\zeta,$$

whence

$$\begin{aligned} \pi n/l_0 &= \pi \int_{\frac{1}{2}}^{\zeta} (1/\sin \pi\zeta) d\zeta \\ &= \ln \tan (\pi\zeta/2). \end{aligned}$$

This can be written

$$\zeta = (2/\pi) \arctan e^{\pi n/l_0}, \quad (6)$$

where $n = 0$ when $\zeta = \frac{1}{2}$, and

$$d\zeta/dn = (1/l_0) \sin \pi\zeta. \quad (6a)$$

Figure 2 gives a graphical illustration of (6). It represents a single dislocation infinitely far from any other in the chain. From (6a) and figure 2 one sees that approximately in a distance corresponding to l_0 atoms the displacement ζ changes by 1. In a relatively small region l_0 , therefore, there exists a state of misfit between the atoms and the substrate, while everywhere else there is a nearly perfect degree of fit. In a positive dislocation there is one more and in a negative dislocation there

is one less atom than there are troughs of the substrate potential. l_0 , which is a measure of the magnitude of the region over which misfit extends, may be called the effective length of a dislocation.

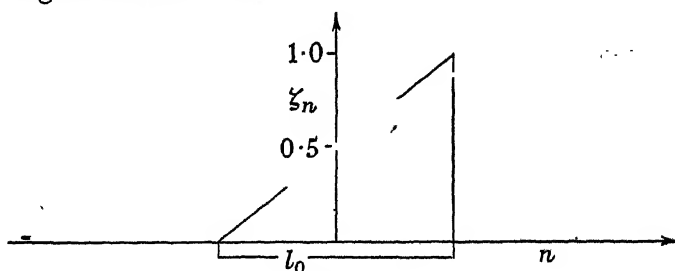


FIGURE 2

(ii) *General solution: $k < 1$*

From (5) it follows that

$$\begin{aligned} \pi n / l_0 k &= \int_0^\phi (1 - k^2 \sin^2 \psi)^{-\frac{1}{2}} d\psi \\ &= F(k, \phi) = F(k, \pi \zeta - \tfrac{1}{2} \pi) \end{aligned} \quad (7)$$

in the standard notation for elliptic integrals. This can be written

$$\zeta = \tfrac{1}{2} + (1/\pi) \operatorname{am}(\pi n / l_0 k)$$

and

$$\begin{aligned} d\zeta/dn &= (1/l_0 k) \operatorname{dn}(\pi n / l_0 k) \\ &= (1/l_0 k) (1 - k^2 \cos^2 \pi \zeta)^{\frac{1}{2}}, \end{aligned} \quad (8)$$

employing the notation of elliptic functions (Jahnke Emde 1938).

Figure 3 illustrates a typical example.

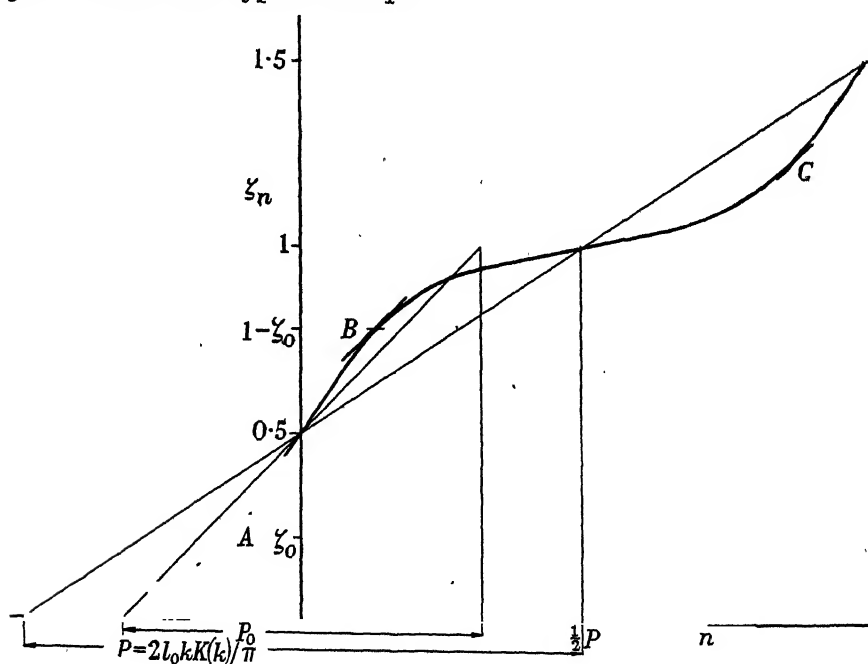


FIGURE 3

We now have a regular sequence of similar dislocations.

From (8) and figure 3 one sees that the effective length of a dislocation is now reduced to kl_0 , and the number of atoms per dislocation is

$$P = 2l_0 kK(k)/\pi, \quad (9)$$

where $K(k)$ is the complete elliptic integral $F(k, \frac{1}{2}\pi)$.

Referring back to figure 1, it is to be observed that even for small values of l_0 , so long as l_0 exceeds unity, there is good qualitative correspondence between solutions of the difference equation (2) and the differential equation (3).

4. EQUILIBRIUM WITH FREE ENDS

On differentiating (1) the term in P_0 , and hence all reference to the natural spring-length b , disappeared. This is only significant in connexion with the boundary conditions. We consider the case in which the chain has a free end whereas the substrate continues.

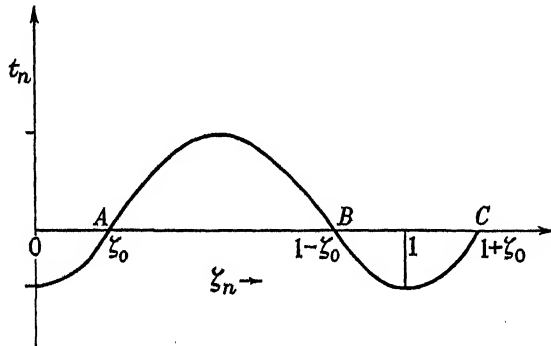


FIGURE 4

For simplicity we speak of tension, which is appropriate for negative dislocations ($l_0 > 0$). The results for positive dislocations are identical.

The tensions in the springs are given by

$$\begin{aligned} t_n &= \mu(x_{n+1} + a - x_n - b) \\ &= \mu a(\zeta_{n+1} - \zeta_n - 1/P_0) \\ &= \mu a \{ d\zeta/dn + (1/2!) d^2\zeta/dn^2 + \dots - 1/P_0 \}. \end{aligned}$$

Neglecting higher terms in the Taylor series, and substituting for $d\zeta/dn$ from (8), we find

$$t_n = (2Wl_0^2/a) \{ (1 - k^2 \cos^2 \pi \zeta)^{1/2} / kl_0 - 1/P_0 \}, \quad (10)$$

and in the special case where $\epsilon = 0$, $k = 1$, $P = \infty$,

$$t_n = (2Wl_0^2/a) \{ (1/l_0) \sin \pi \zeta - 1/P_0 \}. \quad (10a)$$

A graphical representation of (10) is shown in figure 4.

If the springs are cut at the points represented by A , B and C in figures 3 and 4, then the separate parts of the chain will remain in equilibrium with their ends free

—cutting it at B , for instance, will leave that part to the right to B in stable and that to the left in unstable equilibrium. The displacements ζ_0 of the free ends ($t_n(\zeta_0) = 0$) follows from (10), i.e.

$$\cos \pi \zeta_0 = \pm (1/k^2 - l_0^2/P_0^2)^{\frac{1}{2}}, \quad (11)$$

and from (10a), at a free end far from any dislocation,

$$\sin \pi \zeta_0 = l_0/P_0. \quad (11a)$$

5. ENERGIES

5.1. Potential energy per dislocation V_D

From (1) the potential energy per dislocation is

$$\begin{aligned} V_D &= W l_0^2 \sum_{n=0}^{P-1} (\zeta_{n+1} - \zeta_n - 1/P_0)^2 + \frac{1}{2} W \sum_{n=0}^{P-1} (1 - \cos 2\pi \zeta_n) \\ &= W l_0^2 \int_0^P (d\zeta/dn - 1/P_0)^2 dn + W \int_0^P \sin^2 \pi \zeta dn, \end{aligned}$$

where ζ_{n+1} is expanded in a Taylor series, terms of higher order than the first are neglected, and the summation replaced by an integral. Hence, using (7), (8) and (9), and carrying out the integration, this gives

$$V_D = W l_0^2 \{4E(k)/\pi k l_0 - 2(1 - k^2)K(k)/\pi k l_0 - 2/P_0 + P/P_0^2\}, \quad (12)$$

where $E(k) = \int_0^{\frac{1}{2}\pi} (1 - k^2 \sin^2 \psi)^{\frac{1}{2}} d\psi$ is the complete elliptic integral of the second kind.

V_D , thus defined, includes strain energy which is present in the chain (unless $1/P_0$ is zero) in the absence of dislocations. This energy is represented by the final term in (12). We may omit this term if we prefer to regard the undislocated state as the zero of potential energy.

5.2. Generation of a dislocation by force on the free end

A dislocation can be generated by pulling the end atom reversibly from its equilibrium displacement ζ_0 (A , figures 3 and 4) to its next stable equilibrium displacement $1 + \zeta_0$ (C , figures 3 and 4). The work done during this process is

$$\begin{aligned} W_D &= a \int_{\zeta_0}^{1+\zeta_0} t d\zeta \\ &= a \int_0^1 t d\zeta \quad \text{as follows from figure 4} \\ &= 2W l_0^2 \{2E(k)/\pi k l_0 - 1/P_0\}. \end{aligned} \quad (13)$$

The corresponding value for single dislocations then becomes (limit $k = 1$)

$$W_D = 2W l_0^2 (2/\pi l_0 - 1/P_0). \quad (13a)$$

It may be noted that if Q is the total potential energy of a long chain of atoms containing N dislocations, V_D is the mean potential energy per dislocation, Q/N , while W_D is the differential energy per dislocation, dQ/dN . This relationship may be verified by integration.

5.3. Activation energies

Consideration of this process, the application of force to the end atom of the chain, enables us to define an activation energy for the introduction of a dislocation, i.e. the work done in pulling the end atom reversibly from its stable equilibrium displacement (*A*, figures 3 and 4) to its unstable equilibrium displacement (*B*, figures 3 and 4) given by

$$\begin{aligned} U_g &= a \int_{\zeta_0}^{1-\zeta_0} t d\zeta \\ &= 4Wl_0^2 \{E(k, \tfrac{1}{2}\pi - \pi\zeta_0)/\pi kl_0 - (\tfrac{1}{2} - \zeta_0)/P_0\}, \end{aligned} \quad (14)$$

where ζ_0 is given by (11). $E(k, \tfrac{1}{2}\pi - \pi\zeta_0)$ is the incomplete elliptic integral of the second kind.

For dislocations far apart (limit $k = 1$)

$$U_g = 4Wl_0^2 \{(1 - \cos \pi\zeta_0)/\pi l_0 - (\tfrac{1}{2} - \zeta_0)/P_0\}. \quad (14a)$$

Similarly, the activation energy for the escape of dislocations is

$$\begin{aligned} U_e &= a \int_{1+\zeta_0}^{1-\zeta_0} t d\zeta \\ &= U_g - W_D. \end{aligned} \quad (15)$$

6. STABILITY LIMITS

The necessary and sufficient condition that a given solution of equation (3) can correspond to equilibrium when the ends of the chain of atoms are free is that there shall exist a real solution ζ_0 of the equation $t(\zeta_0) = 0$, i.e. equation (11). Hence

$$1 - k^2 l_0^2 / P_0^2 \leq k^2 \quad \text{and} \quad 1 - k^2 l_0^2 / P_0^2 \geq 0,$$

$$\text{i.e.} \quad (1/k^2 - 1)^{\frac{1}{2}} \leq l_0/P_0 \leq 1/k. \quad (16)$$

If $1/k < l_0/P_0$ dislocations will be spontaneously generated at the ends, while if $(1/k^2 - 1)^{\frac{1}{2}} > l_0/P_0$ they will spontaneously escape. When dislocations are absent or far apart (16) reduces to

$$0 \leq l_0/P_0 \leq 1. \quad (16a)$$

The state of lowest energy will be that for which the energy remains stationary for generation or escape of dislocations, i.e. that for which $W_D = 0$, so that

$$l_0/P_0 = 2E(k)/\pi k, \quad (17)$$

and the corresponding equation for dislocations far apart is then

$$l_0/P_0 = 2/\pi. \quad (17a)$$

This defines the critical value of misfit, above which the lowest energy state of the system is one containing dislocations. A somewhat larger misfit, as given by (16a), is needed for them to be developed spontaneously (at the absolute zero of temperature).

7. APPLICATIONS OF THE MODEL

7.1. Imagine a solid crystal with a simple cubic lattice containing one or more straight parallel Taylor (1934) dislocations (all positive or all negative) in a single glide-plane. Suppose the material to be shaved away, parallel to this glide-plane, till there remains only one layer (or a very small number of layers) of atoms above the glide-plane, while the crystal is still thick below it. In the new elastic equilibrium nearly all the strain associated with the dislocation(s) will be in the thin layer above the glide-plane, and the crystal below can be regarded as effectively rigid. Then in each row of atoms immediately above the glide-plane and normal to the dislocation line(s) the conditions are substantially those of our first model, and the equations developed may be applied to the system with reasonably good accuracy. The system may be treated one-dimensionally so long as the dislocation lines remain straight and parallel. In this case b is approximately equal to a , but not exactly, since in general a surface layer of atoms in a crystal has a tendency to expand slightly if held together by homopolar forces, or to contract if they are ionic. Larger differences between b and a arise when we are considering overgrowth or absorption of one substance on a crystal of another.

7.2. Along the edge of an outer layer, or at a place where it is incomplete, ending in a straight terrace edge, we have a row of atoms attracting and repelling each other so as to tend to maintain a natural spacing b and subject to a substantially sinusoidal potential field of wave-length a (not in general exactly equal to b) arising from their neighbours in the rest of the crystal. This is a truly one-dimensional problem, corresponding closely to our first model. We estimate the physical parameters of the model for correspondence with this case from a Lennard-Jones law of forces (1924) between the atoms in § 8, and find that the appropriate value of l_0 , the effective spread of the dislocation, is about 7.

Values of l_0 of similar order of magnitude will apply in case (7.1) when we consider dislocations between a surface monolayer and the remainder of the crystal.

7.3. If in case (7.1) we start with Burgers (1939) 'screw' dislocations in place of the Taylor dislocations, we shall finish with a situation which is again well described by our equations, in which, however, the displacements ζ_n are transverse to the chain of atoms enumerated by n . The atomic positions will in fact correspond directly to the graphs of ζ_n versus n shown in figures 2 and 3. These dislocations, corresponding to screw dislocations in three dimensions, may be called 'transverse': those of case (7.1) corresponding to Taylor dislocations in three dimensions are 'longitudinal'. The longitudinal and transverse dislocations of a surface layer can be of importance in connexion with adsorbed monolayers, oxide layers, and so on. These, and the longitudinal edge dislocations (7.2) all deserve consideration in connexion with the original formation of dislocations in the growth of a crystal.

7.4. A long-chain molecule, say of a normal aliphatic compound, which has some torsional flexibility, and is restrained in twisting by the forces from parallel neighbour chains in the crystal, provides another system to which the same equations are applicable, wherein the displacements ζ_n are angles of rotation.

7.5. A somewhat different physical application is shown in figure 5 which gives a formal two-dimensional description of forced twinning (e.g. as observed in calcite). For the case pictured, an unstrained (two-dimensional) unit cell is supposed to be a 60° rhombus. The state shown, with three or four badly strained unit cells near the centre and around that a rapid asymptotic approach to a strain-free condition, should by its motion translate the twinning surface, causing one twin to grow into the other: and an applied stress should cause it to move. This configuration, however, is not a simple dislocation; its presence does not disturb the continuity of lattice planes. In so far as the twinning plane may be looked upon as a close grid of parallel Taylor dislocations with displacement vectors normal to the plane (Wooster 1940) (which exert forces on each other to hold their neighbours in the same plane, and

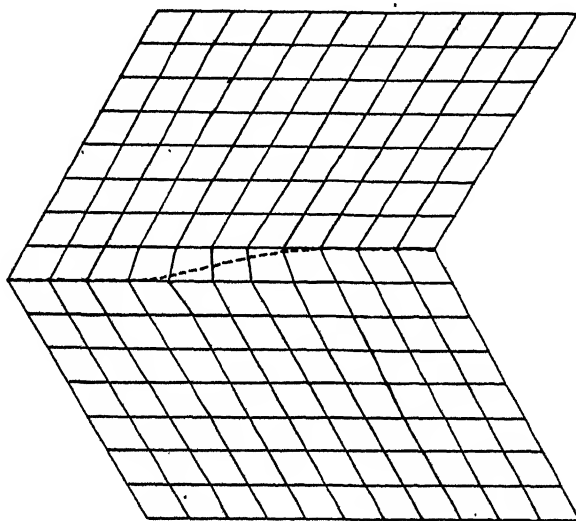


FIGURE 5

are moreover subject to a periodic 'substrate potential' because they are in a crystal), the figure represents a transverse dislocation of the system of dislocations. We may call it a dislocation of second order. Qualitatively, we may describe it by the equations of a one-dimensional dislocatable system developed above, ζ_n representing the displacement of an individual Taylor dislocation. However, at a distance from this second-order dislocation there exists a state of strain similar to that around a single Taylor dislocation; there are $(n+1)$ unit cells in the same distance, measured parallel to the twinning plane, above it, as contains n below it. On this account these second-order dislocations exert long-range repulsive forces on each other. This agrees with observation; Garber's (1947) elastic twin wedges are only about 1μ thick for a length of the order 1 cm., so that $1/P$, the inclination of the twinning surface to the ideal twinning plane, probably does not exceed about $1/10,000$ to $1/1000$.

There are other second-order dislocations, a sigmoid bend of an ordinary dislocation line from one lattice line on to the next parallel one being the simplest type. These are required to describe the inclined twinning surfaces in a Reusch 'rectangle'

or Garber 'leaf' elastic twin of calcite, as contrasted with the Garber wedges. Being dislocations of a one-dimensional system, they also are qualitatively illustrated by the equations investigated.

8. In order to apply the equations obtained for our model to actual physical systems, it is necessary to assign values to the physical parameters, l_0 and P_0 . The most important applications are to a surface 'island' layer of atoms on a crystal, in which we may have dislocations of the edge row with respect to the rest of the island (§ 7.2) or of the layer with respect to the substrate crystal (§ 7.1). In particular, in the latter case the atoms of the surface layer need not be of the same kind as those of the substrate. In this case P_0 can be estimated approximately from the lattice spacings of the bulk materials. To this approximation it becomes infinite when the atoms are of the same kind, but this is not correct. The values both of P_0 and of l_0 may be estimated in this case by the methods of Lennard-Jones. Let the suffixes $X = \text{I, II, III}$, refer to close-packed one-dimensional, two-dimensional, and three-dimensional lattices (the latter in cubic close packing). Then the potential energy per atom may be written

$$V = -A^m S_X \alpha^{-m} + B^n S_X \alpha^{-n},$$

where A and B are constants, α the interatomic distance, and $^m S_X$ signifies the lattice sum for the exponent $-m$ and the lattice X . We shall actually take n to be 12 and m to be 6.

The equilibrium spacing α_X is defined by the condition $\partial V / \partial \alpha = 0$. Thus

$$\alpha_X = (n B^n S_X / m A^m S_X)^{1/(n-m)}.$$

The latent heat per atom, L_X , is minus the potential energy per atom at equilibrium spacing. Thus

$$L_X = A \alpha_X^{-m} S_X (n-m) / n.$$

The force constant in homogeneous compression is

$$\begin{aligned} \mu_X &= (\partial^2 V / \partial \alpha^2)_{\alpha=\alpha_X} = m(n-m) A^m S_X \alpha_X^{-m-2} \\ &= mn L_X \alpha_X^{-2}. \end{aligned}$$

The lattice sums required are

$$\begin{aligned} {}^6 S_{\text{I}} &= 1.017_3, & {}^6 S_{\text{II}} &= 3 \times 1.063_0, & {}^6 S_{\text{III}} &= 6 \times 1.20433, \\ {}^{12} S_{\text{I}} &= 1.0002_5, & {}^{12} S_{\text{II}} &= 3 \times 1.001_6, & {}^{12} S_{\text{III}} &= 6 \times 1.01097, \end{aligned}$$

the values for ${}^6 S_{\text{III}}$ and ${}^{12} S_{\text{III}}$ being those of Jones & Ingham (1925).

If $\alpha_0 = (2B/A)^{1/n}$ represents the equilibrium spacing of a pair (or triangle or tetrahedron) of atoms, we have

$$\begin{aligned} \alpha_0 / \alpha_{\text{I}} &= 1.002_7, & \alpha_0 / \alpha_{\text{II}} &= 1.01_0, & \alpha_0 / \alpha_{\text{III}} &= 1.029, \\ \alpha_{\text{I}} / \alpha_{\text{II}} &= 1.007, & \alpha_{\text{I}} / \alpha_{\text{III}} &= 1.026, & \alpha_{\text{II}} / \alpha_{\text{III}} &= 1.019. \end{aligned}$$

Accordingly, the natural misfit of a row at the edge of an island layer is given by its vernier period

$$P_0(\text{I, II}) = 144,$$

and correspondingly the natural misfit of a surface layer on the three-dimensional crystal is given by

$$P_0(\text{II, III}) = 53.5.$$

To estimate l_0 for the case of the edge row of a two-dimensional lattice we consider the potential energy of a single atom rolling in contact with the 'straight' edge of a semi-infinite two-dimensional lattice. We take the lattice spacing to be uniform, as though the atoms were hard spheres of diameter α_{II} , and the path of the additional atom to be the same as it would be if it were a hard sphere of the same diameter. Then the highest and lowest energies are obtained at the two symmetrical positions where it makes respectively one and two contacts. These energies are

$$1.019A\alpha_{\text{II}}^{-6} = 0.639L_{\text{II}} = 0.274L_{\text{III}},$$

$$0.784A\alpha_{\text{II}}^{-6} = 0.492L_{\text{II}} = 0.193L_{\text{III}},$$

respectively. Taking the fluctuation to be sinusoidal as an approximation (which we shall refine in a later paper in this series), the appropriate value of W in equation (1) is the difference between these two energies:

$$W_{\text{II}} = 0.325A\alpha_{\text{II}}^{-6} = 0.204L_{\text{II}} = 0.0807L_{\text{III}}.$$

We likewise have the force constant

$$\mu_{\text{I}} = 36A\alpha_{\text{I}}^{-8} \times 1.0173.$$

And hence

$$l_0^2 = \mu_{\text{I}}\alpha_{\text{II}}^2/2W_{\text{II}} = 36 \times 1.0173\alpha_{\text{II}}^8/2 \times 0.325\alpha_{\text{I}}^8.$$

Thus

$$l_0(\text{I, II}) = 7.35.$$

l_0 is naturally larger the harder the atoms. Since W is approximately proportional to L , l_0 is approximately proportional to $(mn)^{\frac{1}{2}}$.

We may now find the energy of such an edge dislocation, from equation (13a). It is

$$\begin{aligned} W_D &= W(4l_0/\pi - 2l_0^2/P_0) \\ &= 0.0807L_{\text{III}}(9.37 - 0.75) \\ &= 0.688L_{\text{III}}. \end{aligned}$$

Thus it has just the same order of magnitude as the latent heat of evaporation of a single atom. Such edge dislocations should therefore be fairly frequently formed by thermal agitation.

The value of l_0 should be approximately the same for dislocations of the surface layer with respect to the underlying three-dimensional crystal; for the force constant μ should be about the same in both cases (we are not dealing with homogeneous compression in this second case), and the value of W corresponds to the breaking of one atomic contact per atom in both cases. In this case the dislocation is a line and its energy proportional to the line length of the order of magnitude of an atomic latent heat per atomic spacing of length. It is doubtful whether any appreciable number of dislocations will be formed in thermal equilibrium in this case, but it may be noted that the misfit parameter $l_0/P_0 \approx 0.15$ is sufficiently large to give an activation energy hindering the escape of any dislocations which are formed, except when they are very close together (about 15 atomic spacings apart according to equations (16) and (9)).

We have to thank Professor N. F. Mott for suggesting this problem, and one of us (J. H. van der Merwe) has to thank the South African Council for Scientific and Industrial Research for a grant, and special leave which rendered it possible to perform this research.

REFERENCES

- Burgers, J. M. 1939 *Proc. Acad. Sci. Amst.* **42**, 293.
 Dehlinger, U. 1939 *Ann. Phys., Lpz.*, (5), **2**, 749.
 Frenkel, J. & Kontorowa, T. 1938 *Phys. Z. Sowjet.* **13**, 1.
 Garber, R. 1947 *Russian J. Phys.* **11**, 55.
 Jahnke Emde 1938 *Functionentafeln*, pp. 52–90.
 Jones, J. E. 1924 *Proc. Roy. Soc. A*, **106**, 463.
 Jones, J. E. & Ingham, R. E. 1925 *Proc. Roy. Soc. A*, **107**, 636.
 Lennard-Jones, J. E. 1940 *Proc. Phys. Soc.* **52**, 38.
 Taylor, G. I. 1934 *Proc. Roy. Soc. A*, **145**, 362.
 Wooster, W. A. 1940 *Proc. Phys. Soc.* **52**, 55.

One-dimensional dislocations. II. Misfitting monolayers and oriented overgrowth

BY F. C. FRANK AND J. H. VAN DER MERWE

H. H. Wills Physical Laboratory, University of Bristol

(Communicated by N. F. Mott, F.R.S.—Received 22 December 1948—

Revised 25 March 1949—Read 19 May 1949)

The equations derived in part I of this series for a one-dimensional dislocation model are applied in this paper to the case of a monolayer on the surface of a crystalline substrate, particularly when the natural lattice spacing of the monolayer differs from that of the substrate. Justification is given for this extension of the equations to the two-dimensional case. It is shown that the theory predicts a certain critical amount of misfit (9 % difference in lattice spacing in a simple case) below which the monolayer in its lowest energy state is deformed into exact fit with the substrate, and above which it is only slightly deformed in the mean, having many dislocations between it and the substrate. The energy of adsorption as a function of misfit is also calculated, becoming almost constant above the critical limit. Up to a larger critical misfit (about 14 % in the same simple case) the monolayer can be deposited metastably in exact fit on the substrate, at sufficiently low temperature. Since the dislocated layer is mobile on the surface, completely oriented overgrowth of one crystal on another can only be expected if the first monolayer can be formed over the complete surface under sub-critical conditions. This is in general agreement with observation.

1. PRELIMINARY INVESTIGATION

We propose to apply the equations derived in part I for one-dimensional dislocations to misfitting monolayers on crystalline substrates. Before doing so we examine a little more closely the validity of these equations for a two-dimensional problem. Let us take as a simple first-order representation of the potential energy of a two-dimensional square array of atoms, enumerated by n in the x direction, m in the

y direction, on a crystal face on which the minima of potential energy for single atoms likewise make a square array (e.g. (100) faces of simple cubic or close-packed cubic lattices):

$$\begin{aligned}
 V = & \sum_{n,m} \frac{1}{2} W [1 - \cos 2\pi \zeta_{n,m} + 1 - \cos 2\pi \eta_{n,m}] \\
 & + \sum_{n,m} \frac{1}{2} \mu a^2 [(\zeta_{n+1,m} - \zeta_{n,m} - 1/P_0)^2 + (\eta_{n,m+1} - \eta_{n,m} - 1/P_0)^2] \\
 & + \sum_{n,m} \frac{1}{2} \nu a^2 (\zeta_{n,m+1} - \zeta_{n,m} + \eta_{n+1,m} - \eta_{n,m})^2 \\
 & + \sum_{n,m} \lambda a^2 (\zeta_{n+1,m} - \zeta_{n,m} - 1/P_0) (\eta_{n,m+1} - \eta_{n,m} - 1/P_0). \quad (1)
 \end{aligned}$$

Here $a(n + \zeta_{n,m}) = x_{n,m}$ and $a(m + \eta_{n,m}) = y_{n,m}$ are the co-ordinates of the atom enumerated (n, m) . The first summation expresses the substrate potential, being (for different axes in the two cases) the commencement of a two-dimensional Fourier series representation of the field at the surface of either of the two crystals mentioned above. The second summation expresses the elastic resistance of the monolayer to change of lattice spacing from its natural value b , not necessarily equal to that a of the substrate, as indicated by the vernier period of misfit, $P_0 \doteq a/(b-a)$. The third summation expresses the elastic resistance of the monolayer to shear, the expression in brackets being (for small differences of the displacements) equal to the amount by which the angle made by atoms $(n+1, m)$, (n, m) , $(n, m+1)$ differs from its natural value $\frac{1}{2}\pi$. Thus μ and ν are respectively force constants for compression and shear of the monolayer. λ is a force constant for shear at 45° to the former one.

The conditions for equilibrium of the atom (n, m) are:

$$\begin{aligned}
 0 &= \partial V / \partial \zeta_{n,m} \\
 &= \pi W \sin 2\pi \zeta_{n,m} - \mu a^2 (\zeta_{n-1,m} - 2\zeta_{n,m} + \zeta_{n+1,m}) \\
 &\quad - \nu a^2 (\zeta_{n,m-1} - 2\zeta_{n,m} + \zeta_{n,m+1}) - \nu a^2 (\eta_{n,m-1} - \eta_{n,m} - \eta_{n+1,m-1} + \eta_{n+1,m}) \\
 &\quad - \lambda a^2 (\eta_{n,m+1} - \eta_{n,m} - \eta_{n-1,m+1} + \eta_{n-1,m}) \\
 &\equiv \pi W \sin 2\pi \zeta - \mu a^2 \Delta_n^2 \zeta - \nu a^2 \Delta_m^2 \zeta + \nu a^2 \Delta_{n+\frac{1}{2}} \Delta_{m-\frac{1}{2}} \eta + \lambda a^2 \Delta_{n-\frac{1}{2}} \Delta_{m+\frac{1}{2}} \eta, \quad (2a)
 \end{aligned}$$

$$\text{and} \quad 0 = \pi W \sin 2\pi \eta - \mu a^2 \Delta_m^2 \eta - \nu a^2 \Delta_n^2 \eta + \nu a^2 \Delta_{m+\frac{1}{2}} \Delta_{n-\frac{1}{2}} \zeta + \lambda a^2 \Delta_{m-\frac{1}{2}} \Delta_{n+\frac{1}{2}} \zeta. \quad (2b)$$

Solutions of equations (2a, b) include a number of simple cases. Let us anticipate the existence of solutions in which the cross terms of (2a) are zero. Then (i) if ζ is independent of m , the third term vanishing, or (ii) if ζ is independent of n , the second term vanishing, equation (2a) reduces to the one-dimensional form, being equivalent to equation I (2). In either of these cases the final terms of (2b) vanish, so that (2b) has corresponding simple cases with 'one-dimensional' solutions introducing no cross-term in (2a). Cases (i) and (ii) represent the forms assumed by Taylor 'bridge' and Burgers 'screw' dislocations respectively when they lie in the outermost glide-plane of a crystal. Thus surface dislocations of type (1, 0) (displacement vector $(a, 0)$) and type (0, 1) (displacement vector $(0, a)$) are mutually independent, even when intersecting or superposed, provided that they are of pure Taylor or pure

Burgers form, that is to say, that the dislocation lines run either normal or parallel to their displacement vectors.

In the more general case of a straight dislocation inclined at an angle α to the x -axis, so that

$$\zeta(n, m) = \zeta(p), \quad \eta(n, m) = \eta(p), \quad p = m \cos \alpha - n \sin \alpha,$$

equations (2a, b) reduce, when the differences are approximated as differentials, to

$$0 = \pi W \sin 2\pi\zeta - a^2(\mu \cos^2 \alpha + \nu \sin^2 \alpha) (d^2\zeta/dp^2) - a^2(\nu + \lambda) \sin \alpha \cos \alpha (d^2\eta/dp^2), \quad (3a)$$

and

$$0 = \pi W \sin 2\pi\eta - a^2(\mu \cos^2 \alpha + \nu \sin^2 \alpha) (d^2\eta/dp^2) - a^2(\nu + \lambda) \sin \alpha \cos \alpha (d^2\zeta/dp^2). \quad (3b)$$

Neglecting the final term of (3a), it reduces to the one-dimensional form equivalent to I (3) with

$$l_0 = a\{(\mu \cos^2 \alpha + \nu \sin^2 \alpha)/2W\}^{\frac{1}{2}}.$$

However, unless $(\nu + \lambda) \sin \alpha \cos \alpha$ is zero, the presence of ζ displacements then demands η displacements also, to satisfy (3b), which feed back a final perturbing term into (3a). The perturbing term is very small if $(\nu + \lambda) \sin \alpha \cos \alpha$ is small and there are no (0, 1) dislocations. Thus the one-dimensional equations give a good approximate representation of skew dislocations as well as those of normal form, but skew dislocations of (1, 0) and (0, 1) type interfere with each other where they intersect or overlap.

As a special case (1, 1) dislocations of any inclination reduce simply to the one-dimensional form. For if $\eta = \pm \zeta$ equations (3a) and (3b) both reduce to

$$0 = \pi W \sin 2\pi\zeta - a^2(\mu \cos^2 \alpha + \nu \sin^2 \alpha \pm (\nu + \lambda) \sin \alpha \cos \alpha) (d^2\zeta/dp^2), \quad (4)$$

which is of one-dimensional form with

$$l_0 = a\{(\mu \cos^2 \alpha + \nu \sin^2 \alpha \pm (\nu + \lambda) \sin \alpha \cos \alpha)/2W\}^{\frac{1}{2}},$$

e.g. if $\alpha = \frac{1}{4}\pi$ we have for the (1, 1) Taylor dislocation

$$l_0 = a\{(\mu + 2\nu + \lambda)/4W\}^{\frac{1}{2}},$$

and for the (1, $\bar{1}$) Burgers dislocation

$$l_0 = a\{(\mu - \lambda)/4W\}^{\frac{1}{2}},$$

contrasted with

$$l_0 = a(\mu/2W)^{\frac{1}{2}} \quad \text{and} \quad l_0 = a(\nu/2W)^{\frac{1}{2}}$$

for the (1, 0) and (0, 1) Taylor and Burgers surface dislocations respectively.

We omit further exploration of the mutual interaction of dislocations of different type, having established the general applicability of the one-dimensional equations to the two-dimensional problem, and the mutual independence of crossed dislocations in important cases.

2. INTRODUCTION

Having confirmed the relevance of the one-dimensional dislocation equations to the two-dimensional monolayer, we shall use these equations to calculate the energy and stability of dislocated and non-dislocated states of the monolayer, for various

degrees of misfit. Mathematically, it will be convenient to express the dislocation density indirectly, by means of the parameter k (the argument of the complete elliptic integrals). On this account we may advantageously preface the mathematical treatment by a qualitative explanation. Supposing there are no dislocations between monolayer and substrate, the monolayer therefore being homogeneously deformed to fit exactly on to the substrate, the strain energy involved depends quadratically upon the misfit. On the other hand, if there is no misfit, the formation of dislocations stores an amount of strain energy proportional to their number, whether they are positive or negative, until their density becomes high when there is additional energy due to their interaction. Thus, initially, the energy increases quadratically with the misfit and linearly with the dislocation density. When there is a finite amount of misfit (let us say, the monolayer is in compression) the energy required to form a positive dislocation is increased, and the energy required to form a negative dislocation is reduced. This reduction increases proportionally with the amount of misfit, and at a certain critical amount of misfit the net amount of energy required to make a dislocation becomes zero. Since the dislocations do not interact appreciably until they are close together, the dislocation density in the equilibrium state increases almost abruptly from zero to a large value on passing this critical amount of misfit. Thereafter, the dislocation density approaches asymptotically to the reciprocal of the vernier period of misfit, and the mean lattice spacing of the monolayer approaches asymptotically to its natural, unstrained, spacing. Numerically, the critical amount of misfit is shown to correspond to a difference of 9 % in the lattice spacings for a simple case, though there may be rather wide variation about this value, depending on the relative forces exerted by monolayer atoms on each other and on the substrate respectively. The energy of adsorption varies parabolically with the amount of misfit up to the critical value, and then rapidly approaches a constant asymptotic value.

This is not the only critical condition of importance, for there still remains an activation energy for generation of dislocations, which only falls to zero with a larger degree of misfit (about 14 % in the same simple case). Unless the misfit exceeds this second critical value, it is possible for the monolayer to be deposited, at low temperature, in exact fit on the substrate. This represents a metastable state which could eventually be destroyed by thermal activation.

3. MISFIT-DISLOCATION DENSITY RELATIONSHIPS AND STABILITY

It is clear from the foregoing that the energy and hence stability of a deposited monolayer on a crystalline substrate depends upon the degree of misfit and dislocation density. To every energy state, e.g. the lowest energy state, there corresponds a definite relationship between these two quantities.

Referring back to part I, equation I(9) provides us with an expression for the dislocation density $1/P$ in terms of k :

$$1/P = \pi/2l_0 kK(k). \quad (5)$$

Similarly equation I (12) gives us the expression for the potential energy per dislocation of the deposit

$$V_D = W l_0^2 \{ 4E(k)/\pi l_0 k + 2(k^2 - 1) K(k)/\pi l_0 k - 2/P_0 + P/P_0^2 \}.$$

Hence the mean potential energy per atom of the deposit is

$$V = V_D/P = W \{ 2E(k)/k^2 K(k) + 1 - 1/k^2 - \pi l_0/k K(k) P_0 + l_0^2/P_0^2 \}. \quad (6)$$

For a given monolayer and substrate the misfit is fixed, so that the energy of the system depends further only on the concentration of dislocations, which is measured by $(1/P)$. Since $1/P$ is defined in terms of k it is convenient to make use of the relationship

$$\frac{\partial V}{\partial(1/P)} = -P^2 \frac{\partial V}{\partial k} \frac{\partial k}{\partial P}, \quad \frac{d}{dk} [kK(k)] = \frac{E(k)}{1-k^2} \quad \text{and} \quad \frac{d}{dK} \left[\frac{E(k)}{k} \right] = -\frac{K(k)}{k^2}.$$

Hence

$$\frac{\partial V}{\partial(1/P)} = 2W l_0 \{ 2E(k)/\pi k - l_0/P_0 \}.$$

For values of $l_0/P_0 > 2/\pi$ the lowest energy state is given by $\partial V/\partial(1/P) = 0$, and therefore

$$l_0/P_0 = 2E(k)/\pi k. \quad (7)$$

For lower values of l_0/P_0 , the expression for $\partial V/\partial(1/P)$ is always positive, and the lowest energy state is always given by $1/P = 0$ up to the critical condition $l_0/P_0 = 2/\pi$, corresponding to $k = 1$ (see figure 1).

In part I we have also derived the condition I (16), namely,

$$(1/k^2 - 1)^{1/2} \leq l_0/P_0 \leq 1/k, \quad (8)$$

which determines for a given value of k , and hence given $1/P$, the limits of $1/P_0$, between which the system is in a metastable state, when its boundaries are free. The limiting condition for there being any dislocations at all is $k = 1$, and hence corresponds to

$$0 \leq l_0/P_0 \leq 1.$$

The limit $l_0/P_0 = 1/k$ corresponds to spontaneous dislocation, i.e. the state in which the generation of another dislocation needs no activation energy, whereas the limit $l_0/P_0 = (1/k^2 - 1)^{1/2}$ signifies spontaneous escape of dislocations, i.e. the state in which one of the existing dislocations can escape without activation energy. (For graph see figure 1.)

The graphs refer to negative dislocations ($P_0 > 0$, $l_0 > 0$). The corresponding graphs for positive dislocations ($P_0 < 0$, $l_0 < 0$) are the reflexions of these curves in the l_0/P_0 axis. This symmetrical character arises from our use of Hooke's law for forces between atoms of the monolayer and cannot accurately correspond to reality.

It may also be noted that $1/P = \bar{b}/a - 1$, where \bar{b} is the average lattice spacing of the dislocated system, and $1/P_0 = b/a - 1$, where b is its natural spacing. Thus the graph of l_0/P against l_0/P is likewise a graph of $l_0(\bar{b}/a - 1)$ against $l_0(b/a - 1)$.

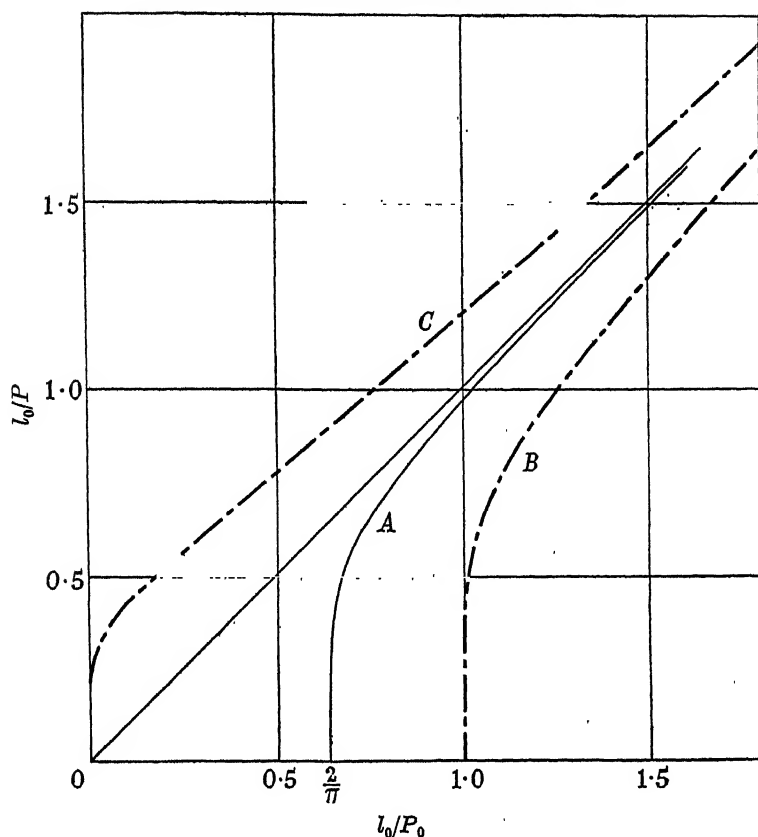


FIGURE 1. Dependence of relative dislocation density l_0/P or $l_0(b/a-1)$ on relative misfit l_0/P_0 or $l_0(b/a-1)$. *A*, lowest energy state (note $l_0/P = 0$ when $0 \leq l_0/P_0 \leq 2/\pi$). *B*, spontaneous dislocation. *C*, spontaneous escape. In an average case in which $l_0 \sim 7$, the critical condition $l_0/P_0 = 2/\pi$ corresponds to a misfit of about 9 %.

4. VARIATION OF MEAN POTENTIAL ENERGY PER ATOM WITH DISLOCATION DENSITY

For a given monolayer and substrate, i.e. for given values of l_0/P_0 , the mean potential energy per atom as a function of relative dislocation density, $V(l_0/P)$, is determined by equations (5) and (6). (See figure 2 for graphs.)

The variation of V against l_0/P , in the state of lowest energy of the system, is naturally one of great importance. The necessary relationship $V_0(l_0/P)$ is easily obtained by elimination of l_0/P_0 between (6) and (7), giving

$$V_0(l_0/P) = W\{4E(k)^2/\pi^2k^2 + 1 - 1/k^2\}. \quad (9)$$

In the limit when there are no dislocations ($k = 1$, $l_0/P_0 = 2/\pi$) this is equal to $4W/\pi^2$.

Elimination of l_0/P_0 between (6) and (8) gives the limiting values of $V(l_0/P)$ for which dislocations are statically persistent. Substitution of $l_0/P_0 = 1/k$ renders the function $V_0(l_0/P)$, which defines the relationship between relative dislocation density l_0/P and the limits of the mean energy per atom beyond which the chain will dislocate spontaneously. Hence

$$V_0(l_0/P) = W\{(2E(k) - \pi)/k^2K(k) + 1\}. \quad (10)$$

Whence it follows that the limiting value of the mean energy per atom, at which a chain containing no dislocations will start to dislocate spontaneously, is given by

$$[V_g(l_0/P)]_{k=1} = W.$$

The limiting misfit concerned is $(1/P_0)_{k=1} = 1/l_0$.

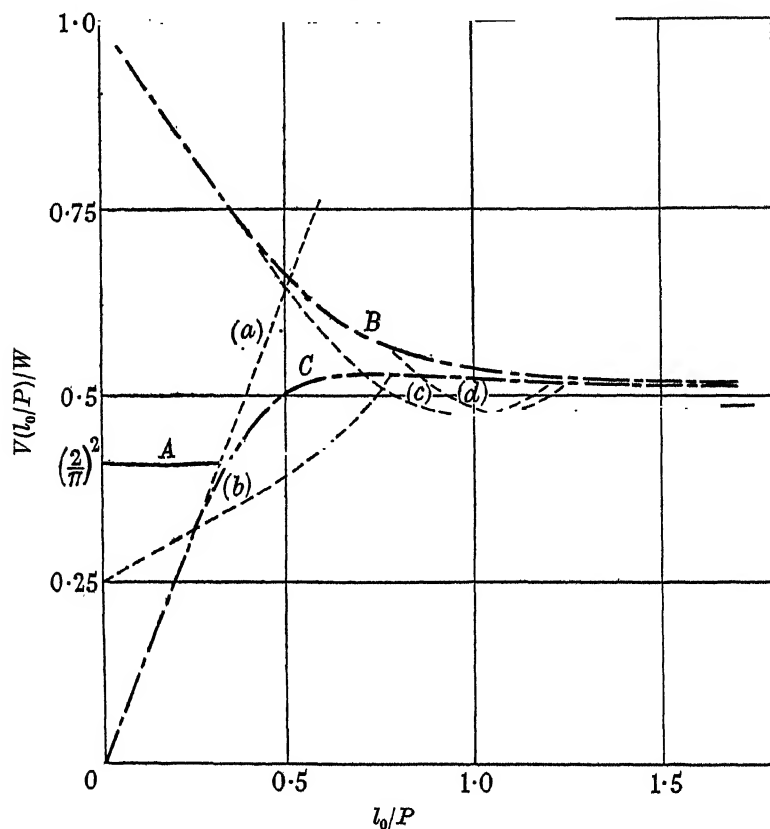


FIGURE 2. Dependence of mean potential energy per atom (in units of W) on relative dislocation density, $V(l_0/P)$ or $V(l_0[b/a-1])$, with (a) $l_0/P_0 = 0$, (b) $l_0/P_0 = 0.5$, (c) $l_0/P_0 = 1.0$, (d) $l_0/P_0 = 1.1$, shown over region of dislocation persistence. A, $V_0(l_0/P)$, lowest energy state; B, $V_g(l_0/P)$, spontaneous dislocation; C, $V_e(l_0/P)$, spontaneous escape.

Similarly, substitution of $l_0/P_0 = (1/k^2 - 1)$ in (6) gives us the corresponding function for the spontaneous escape of dislocations

$$V_e(l_0/P) = W\{2E - \pi(1 - k^2)^{\frac{1}{2}}\}/k^2 K(k). \quad (11)$$

Again, in the limiting case of no dislocations, this is zero, as is l_0/P_0 .

5. VARIATION OF MEAN POTENTIAL ENERGY PER ATOM WITH MISFIT

The functional relationship of mean potential energy per atom with dislocation density becomes clear, when one writes (6) in the form

$$V(l_0/P_0) = W\{(l_0/P_0 - \pi/2kK(k))^2 + 2E(k)/k^2K(k) + 1 - 1/k^2 - \pi^2/4k^2K(k)^2\}, \quad (12)$$

which displays the parabolic relationship of $V(l_0/P_0)$, when $1/P$ and hence k is fixed. Apart from the shift of the vertex

$$(\pi/2kK(k), 2E(k)/k^2K(k) + 1 - 1/k^2 - \pi^2/4k^2K(k)^2)$$

of the parabola by a change in dislocation density, it further remains unaltered, as shown by the graphs in figure 3, where there are also the corresponding graphs for the three most important states of the system.

When $b = \bar{b} = 2a$ say, the mean energy per atom will obviously be zero. Hence there are also ordinates of symmetry at $1/P_0 = 1, 2, \dots$. The mathematical expressions do not reveal this property, because of the approximation of a difference equation by a differential equation.

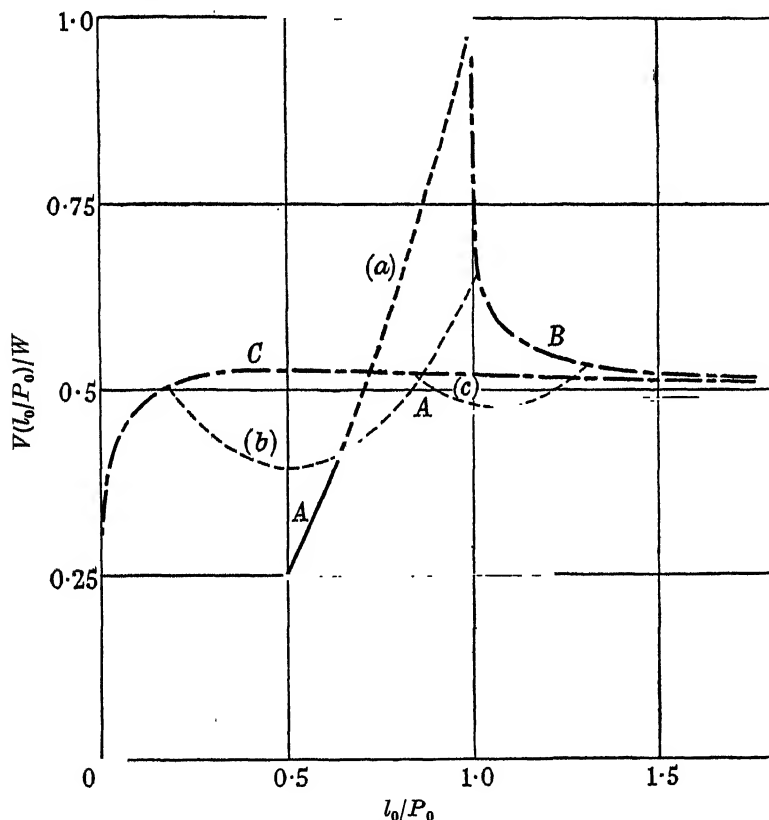


FIGURE 3. Dependence of mean potential energy per atom (in units of W) on relative misfit, $V(l_0/P_0)$ or $V(l_0[b/a-1])$. Parabolas: (a) $l_0/P = 0$, (b) $l_0/P = 0.506$, (c) $l_0/P = 1.059$, shown over region of dislocation persistence. A, $V_0(l_0/P_0)$, lowest energy state, quadratic dependence up to critical point $l_0/P_0 = 2/\pi$, henceforth slow asymptotic increase to $W/2$; B, $V_0(l_0/P_0)$, spontaneous dislocation; C, $V_s(l_0/P_0)$, spontaneous escape. Points on A, B and C determined from corresponding curves in figures 1 and 2.

6. APPLICATIONS TO MISFITTING MONOLAYERS

Application of the foregoing to misfitting monolayers involves knowledge of the physical parameter $l_0 = (\mu a^2/2W)^{1/2}$. An estimate of l_0 was already made in part I, § 8, for a monolayer on a crystalline substrate, both consisting of atoms of the same kind

and assuming Lennard-Jones forces between atoms. The value obtained, namely, $l_0 = 7$, depends on the rigidity of the monolayer and the binding energy between atoms of the monolayer and substrate, and can therefore be appreciably different for different combinations. It will be larger when the atoms of deposit are relatively incompressible, and weakly bound to the substrate, smaller when they are relatively soft and strongly bound.

The degree of misfit $1/P_0$ was defined for our one-dimensional model as $(b-a)/a$, where b is the equilibrium separation of atoms in the chain and a the substrate potential wave-length. In our application of this theory to monolayers on crystalline substrates, b and a will generally be the nearest neighbour distances of atoms of the natural lattices of deposit and substrate in the junction plane.

One may therefore expect that up to the critical misfit $2/\pi l_0 \sim 9\%$, the deposited monolayer will attain the same lattice spacing as the crystalline substrate, and will be free from dislocations and fixed with respect to the substrate. For degrees of misfit exceeding this critical value but less than the limiting value $1/l_0 \sim 14\%$, the generation of dislocations still needs an activation energy, and at sufficiently low temperatures monolayers can still be deposited undergoing deformation to a state of fit with the corresponding substrate.

7. ORIENTED OVERGROWTH

Considerations on misfitting monolayers further lead to predictions for some of the necessary conditions under which the orientations of thin films grown on crystalline substrates are related to those of the substrates. For degrees of misfit less than the critical values obtained in the previous section, the first monolayer that goes down takes up the same lattice spacings as the substrate, contains no dislocations, and is therefore rigidly fixed to the substrate. If the natural lattices of deposit and substrate are isomorphic, the pattern of atoms in this first monolayer is identical with that in a plane of the natural deposit lattice, and hence the next monolayer can grow on to this under conditions similar to those in which the first has grown on to the substrate. Evidently this growth process can also take place for non-isomorphic lattices, provided the lattice of the deposit contains planes whose atomic configuration resembles that of the exposed surface of the substrate: this condition can only be satisfied on specific crystallographic surfaces of the substrate.

One may, however, expect the critical conditions at the free boundaries to change with film thickness, e.g. thickening will certainly cause the generation of dislocations at the free boundaries of an initially undislocated film, since the energy required to compress the thick film is much greater. Thus, a layer of two atoms thick will have a critical misfit of the order of $(2)^{-\frac{1}{2}}$ that of a monolayer, since l_0 is proportional to $\mu^{\frac{1}{2}}$ (taking the elastic constant for a double layer to be twice that of a monolayer). On the other hand, once the first monolayer is completed it does not contain 'free boundaries on a flat substrate', since the film will also be completed around corners and edges and hence the boundaries will be pinned to the substrate. Its roughness further stabilizes the hold of the substrate on the film. A stable oriented film, pseudo-morphic with the substrate, can therefore grow in this way, the essential condition

being that the first monolayer initially fits on to the substrate, has a similar atomic configuration to a plane in the natural deposit lattice, and covers any flat region of the surface completely.

The large strains permissible in the thin films cannot, of course, persist into layers of indefinite thickness, a fact well established by experiments showing that pseudomorphic growth is no longer observed in sufficiently thick films. It will obviously be impossible to grow macroscopically thick layers with more than, say, 0.1 % of strain, corresponding to the yield stress of the bulk material. Thickening of films is therefore accompanied by processes which allow the bulk of thick deposits to be free from the large strains of the initial thin films. This transition from thin to thick films is worth more detailed theoretical and experimental investigation.

The general problem of oriented overgrowth on various surfaces of a substrate non-isomorphic with the deposit is exceedingly complicated, but from similar theoretical considerations to those above, one may anticipate that a two-degree orientation can exist when there is a similarity in spacing in one row of closely packed atoms in each lattice, as remarked by Wilman (1940).

We have already pointed out that one and the same critical misfit is not to be expected for all sorts of substances on all sorts of substrates, but various critical misfits for various classes of case, viz. large critical misfit when there is particularly strong binding between deposit and substrate and low rigidity of deposit, and conversely. For example, the large critical misfit characteristic of the oxides and iodides can be partly accounted for by the greater deformability of oxide and iodide ions respectively.

These theoretical considerations are in good general agreement with the experimental observations which are summarized by van der Merwe (1949).

We should like to thank F. R. N. Nabarro for discussion and constructive criticism.

REFERENCES

- van der Merwe, J. H. 1949 To be submitted to the Faraday Society Discussion on Crystal Growth, April 1949.
Wilman, H. 1940 *Proc. Phys. Soc.* 52, 323.

The rate of evaporation of droplets.

II. The influence of changes of temperature and of the surrounding gas on the rate of evaporation of drops of di-*n*-butyl phthalate

By J. BIRKS AND R. S. BRADLEY

Department of Inorganic and Physical Chemistry, University of Leeds

(Communicated by M. G. Evans, F.R.S.—Received 28 December 1948)

The vapour pressure of di-*n*-butyl phthalate has been determined by Knudsen's method at 15 to 40° C and is given by $\log_{10} p$ (in microns) = $-4790/T + 14.502$. The latent heat of evaporation is 21,910 cal. per mole. Rates of effusion through a small hole in the presence of pressures of air 2 to 0.01 cm. have also been determined and shown to conform to the equation rate of loss of mass = $A/(B+P)$, where A and B are constants. The rate of evaporation of drops of di-*n*-butyl phthalate about 0.5 mm. in radius has been determined at 15 to 40° C, in air at pressures 20 to 0.01 cm. and in hydrogen at 19.90° C and Freon 12 (CCl_2F_2) at 19.90 and 30.00° C and has been accounted for theoretically. Within the experimental error the evaporation coefficient is unity at all temperatures. The diffusion coefficient D in air at 76 cm. varies from 0.0341 to 0.0473 for a temperature variation 15 to 40° C, D being proportional to T^3 . The diffusion coefficient in H_2 at 19.90° C is 0.153 for 76 cm., and for diffusion in Freon 12 at 76 cm. is 0.0126 at 19.90° C and 0.0140 at 30.00° C. The collision radius of di-*n*-butyl phthalate is 4.45 Å as determined from the experiments in air at 19.90° C and the sum of the radii of di-*n*-butyl phthalate and air molecules is proportional to T^{-1} . From the experiments in hydrogen the collision radius of di-*n*-butyl phthalate is 4.68 Å at 19.90° C and from the experiments using Freon 12, 5.24 Å at 19.90° C. The shape of the drops has been studied, and errors introduced by assuming a spherical shape are shown to be negligible.

INTRODUCTION

In part I (Bradley, Evans & Whytlaw-Gray 1946) experiments on the rate of evaporation of drops of di-*n*-butyl phthalate at 20° C were described, and were shown to be adequately accounted for theoretically. This work has been extended to cover the temperature range 15 to 40° C, and, in addition to air, as the diffusion medium hydrogen and Freon (CCl_2F_2) have also been used. In a manner similar to that described in part I the vapour pressures of di-*n*-butyl phthalate were also determined over the same temperature range, with the results described in the next section.

THE VAPOUR PRESSURE OF DI-*N*-BUTYL PHTHALATE AT 15 TO 40° C

The di-*n*-butyl phthalate was shaken with 10 % sodium carbonate solution to remove phthalic acid, and after washing and drying was vacuum-distilled at 215° C; it was then subjected to molecular distillation at less than 10^{-5} cm. of mercury at 90° C. The middle fraction, which was used for evaporation, had a refractive index n_D 1.49292 at 20° C and densities given by the equation

$$\rho'_4 = 1.06199 - 8.15 \times 10^{-4}t',$$

where t' is the centigrade temperature. Other materials used, such as mercury and benzophenone, were also purified.

A differential Knudsen method was used, as described in part I, the loss in weight of tubes containing the liquid, and closed by platinum caps pierced by small holes, being determined over periods of from 5 to 10 hr. In this work and in the droplet evaporation work described later the air thermostat used in part I was replaced by a water thermostat constant to 0.01°C , and fitted with a removable plate-glass front to facilitate manipulations. The tubes were calibrated using mercury and benzophenone, the vapour pressure of which is given by Neumann & Völker (1932); stable solid benzophenone $\log_{10} p = -4966/T + 17.46$, liquid benzophenone $\log_{10} p = -4087/T + 14.75$, p in microns of mercury. An additional check was provided by comparing the results with liquid and solid benzophenone, good agreement between the calibration constants being obtained. The absence of self-cooling was revealed by the agreement between the results for hole sizes 0.0095 to 0.0185 cm.^2 in area; the diameter of the hole was less than the mean free path of vapour molecules, a necessary condition for Knudsen's formula to apply. Results are given in table 1.

TABLE 1. VAPOUR PRESSURE OF DI-*n*-BUTYL PHTHALATE, USING TUBES

temp. ($^\circ\text{C}$)	$p \times 10^5$ (mm.)	$p \times 10^5$ (mm.) smoothed value
19.90	1.43	1.42
23.00	2.05	2.11
25.00	2.73	2.72
30.00	5.0	5.00
35.00	8.85	9.00
40.00	15.8	15.95
43.00	22.0	22.3
43.50	22.2	23.6
44.00	25.1	25.0

The smoothed values were taken from the equation

$$\log_{10} p = -4790/T + 14.502 \quad (p \text{ in microns of mercury}), \quad (1)$$

which gave a latent heat of evaporation of $21,910\text{ cal.} \pm 5\%$ per mole.

The disadvantage of the method described, using a micro-analytical balance, is that it is a slow process, and errors are introduced in timing, since the time when effusion starts on reducing the pressure is uncertain. These errors are reduced by using long times, but weighing errors may arise from condensation of moisture, etc. It was therefore decided to use the quartz microbalance to weigh the effusion apparatus, so that the weight could be continuously recorded over short periods. This necessitated a new design for the effusion vessel, which had to be light enough to be carried by the quartz microbalance.

Pyrex tubing was drawn into a capillary, and a bulb 1 cm. in diameter was blown at the end and two small holes on opposite sides of the bulb were then blown by means of point flames. This effusion pot carrying 0.02 g. of liquid was hung from the quartz microbalance in the neighbourhood of a suitable absorbent, and deflexion-time readings were taken with a hard vacuum in the apparatus. The liquid was

injected into the pot by means of micro-syringe, care being taken not to wet the sides of the holes.

Some difficulty was encountered over the choice of absorbent, tricresyl phosphate being finally used. It will be recalled that in part I charcoal gave good results for the absorption of vapours from droplets down to air pressures of about 0.1 mm. of mercury. It was not found possible, however, to obtain a hard vacuum round a hanging pot using charcoal as absorbent, as was revealed by low effusion rates. This is not surprising considering that the charcoal cage had to be assembled round the hanging pot in the presence of moist air from the empty thermostat. The best results with charcoal were obtained by successive introduction and removal of small quantities of air. Similar difficulties were encountered using silica gel. A surface near the hanging pot and cooled by liquid air was also used, but the cooling had to be started when the system was evacuated, to prevent cooling the pot, and this cooling set up oscillations in the balance. No doubt these could have been removed by electro-magnetic damping, but this was considered an unnecessary elaboration, especially as the hanging pot showed signs of cooling, i.e. lower rates of effusion, even in the hardest vacuum. With liquid air around an external trap the initial effusion rates were in agreement with the results using tricresyl phosphate, but decreased with time, owing possibly to build up of vapour.

A pool of degassed tricresyl phosphate gave reproducible results. This liquid has an extremely low vapour pressure, since a drop hung on the microbalance and subjected to a hard vacuum for 3 hr. at 20° C gave no measurable deflexion, i.e. the vapour pressure was less than 2×10^{-8} mm. of mercury.

The difficulty with hanging pots, which is almost absent with the tubes embedded in metal, is the self-cooling. Since the vacuum round the pots was hard, effusion resulted in continuous cooling (contrast the self-cooling of drops discussed later) and runs were therefore made as short as possible. If Q is the rate of loss of mass, in g./sec., t the time in sec., S_1 the specific heat of the substance studied and W_1 its weight, S_2 and W_2 the corresponding quantities for Pyrex glass, L/M the latent heat of evaporation per g., M the molecular weight, ΔT the self-cooling, then

$$\Delta T = \frac{L}{M} t Q / (S_1 W_1 + S_2 W_2).$$

With di-*n*-butyl phthalate this gives self-cooling corresponding to a decrease in rate of less than 3 %, but with benzophenone errors of 8 % are possible; this implies that the vapour pressures in the most unfavourable case are in error to the extent of 5 %. However, no allowance is made in this calculation for conduction and radiation. Results are given in table 2.

Owing to self-cooling the hanging pot is less reliable than the tube at high temperature, but the reverse is true of the lower temperatures. The final equation for vapour pressure is that given above.

Our results for the vapour pressure and latent heat of di-*n*-butyl phthalate may be compared with the published values given in table 3. The results for di-*n*-butyl phthalate published in part I (3.05×10^{-5} mm. at 20° C) are high, possibly owing to impurity; we believe equation (1) to be accurate to 5 %.

TABLE 2. VAPOUR PRESSURES OF DI-*n*-BUTYL PHTHALATE
IN 10⁻⁵ MM. OF MERCURY

temp. (°C)	pot 4 hole area 0.751 mm. ²	pot 5 hole area 0.48 mm. ²	differential Knudsen method (smoothed)
15.00	0.73	0.75	0.75 (extrapolated)
19.90	1.31	1.38	1.42
25.00	2.46	2.59	2.72
30.00	4.84	4.91	5.00
35.00	8.7	8.53	9.00
40.00	—	14.4	16.00

TABLE 3. PUBLISHED VAPOUR PRESSURES OF DI-*n*-BUTYL PHTHALATE

author	vapour pressure (mm. × 10 ⁵)	temp. (°C)	latent heat of evaporation	method
Zabel (1933)	4.2	23 ± 2	—	ionization gauge
Woodland & Mack (1933)	0.31	25	—	Knudsen effusion
Hickman (1936)	1.9	25	—	ionization gauge
Hickman (1937)	3.27	25	21,400	extrapolated from 50 to 100°C using hanging disk
Verhoek & Marshall (1939)	1.29	25	23,400	do.
Burrows (1946)	1.9	25	—	extrapolated from 150°C
Kapff & Jacobs (1947)	1.0	25	25,500	molecular dew apparatus
this paper	2.72	25	21,910	Knudsen effusion

When the solid absorbents were used some concern was felt as to the attainment of high vacuum in the neighbourhood of the hanging pot. A study was therefore made on the influence of air pressure on effusion rate, in order to determine whether the rates of effusion in a hard vacuum obtained by extrapolation from high pressures agreed with the experimental hard-vacuum rates. So far as we know the influence of air pressure has not been previously studied, and because of their intrinsic interest our results are briefly reported in the next section, although they were not used in calculations of the vapour pressure.

RATES OF EFFUSION FROM A HANGING POT IN THE PRESENCE OF AIR

Two pots were used, of hole areas 2.11 mm.² (pot 2) and 0.751 mm.² (pot 4) and rates of effusion studied at air pressures of 2 to 0.01 cm., using di-*n*-butyl phthalate and benzophenone. In all cases the rates of loss of mass Q was found to be expressed by an equation of the form

$$Q = \frac{A}{P(B/P + 1)}, \quad (2)$$

the vacuum rate of evaporation being A/B .

Table 4 gives the results with tricresyl phosphate as absorbent, using pot 2, for solid benzophenone at 19.9° C. Each value of Q was itself derived from a number of points on the line for deflexion of microbalance versus time. By plotting Q_0/Q

versus $1/P$ (where Q_0 is the rate A/P , derived from the first part of the Q versus $1/P$ curve, which is nearly linear) the constant B is easily derived, giving

$$Q = \frac{0.16 \times 10^{-6}}{P(0.0085/P + 1)} \text{ g.min.}^{-1} \quad (P \text{ in cm. of mercury}).$$

TABLE 4. EFFUSION OF BENZOPHENONE AT A RANGE OF AIR PRESSURES

P (cm. Hg)	$Q \times 10^6 \text{ g.min.}^{-1}$	$Q \times 10^6 \text{ g.min.}^{-1}$ (calc.)
2.120*	0.0775	0.0753
1.365	0.125	0.116
1.002*	0.151	0.158
0.601	0.270	0.263
0.564*	0.265	0.280
0.288*	0.514	0.540
0.282	0.516	0.551
0.0740	1.95	1.94
0.0507	2.66	2.70
0.0337	3.81	3.80
0.0226	5.16	5.15
0	—	18.8

* Charcoal as absorbent.

It should be noted that there is no disadvantage to the use of charcoal as absorbent in this region of air pressure; indeed, it functions as well as for the droplet runs.

For pot 2, using di-*n*-butyl phthalate at 19.9° C,

$$Q = \frac{8.15 \times 10^{-9}}{P(0.0080/P + 1)} \text{ g.min.}^{-1} \quad (P \text{ in cm. of mercury}).$$

For pot 4 the results were:

$$\text{Solid benzophenone } Q = \frac{8.80 \times 10^{-8}}{P(0.0140/P + 1)} \text{ g.min.}^{-1} \quad (P \text{ in cm. of mercury}).$$

$$\text{Di-}n\text{-butyl phthalate } Q = \frac{5.58 \times 10^{-9}}{P(0.0140/P + 1)} \text{ g.min.}^{-1} \quad (P \text{ in cm. of mercury}).$$

These equations enable us to calculate the rate of evaporation *in vacuo*, whence the vapour pressure may be derived; for pot 2 the vapour pressure of di-*n*-butyl phthalate was 1.52×10^{-5} mm., for pot 4, 1.67×10^{-5} mm., in reasonable agreement with the results quoted in the previous section, 1.42×10^{-5} mm. at 19.9° C.

Results suggested that conical diffusion from the hole was occurring, the vapour spreading out fan-wise. An exact treatment is intractable, but an approximate analysis on lines similar to those used in part I (where, however, the treatment was exact) is given below.

Suppose that the concentration of vapour in the pot with two holes of radius a is c_0 , and that c_1 is the concentration of vapour just outside the hole. Let the half-angle of the cone be a/r . The rate of loss mass by conical diffusion from the region just outside the hole is given by

$$-\frac{dm}{dt} = \frac{\pi a^2}{r^2} \frac{m^2 D c_1}{(1/r - 1/r_0)},$$

where D is the diffusion coefficient at the air pressure used, m_2 is the mass of diffusing molecule, and r_0 the distance from hole to absorbent. The rate of transfer of mass across the hole is

$$-\frac{dm}{dt} = \pi a^2 m_2 \nu (c_0 - c_1),$$

where $\nu = \left(\frac{kT}{2\pi m_2}\right)^{\frac{1}{2}}$. Equating these values gives the value of c_1 , whence for two holes,

$$-\frac{dm}{dt} = \frac{2\pi a D m_2 c_0}{\frac{D}{a\nu} + \frac{r}{a} \left(1 - \frac{r}{r_0}\right)},$$

or for $r_0 \gg r$,

$$-\frac{dm}{dt} = \frac{2\pi a D m_2 c_0 (a/r)}{\frac{D}{a\nu} \left(\frac{a}{r}\right) + 1}. \quad (3)$$

Results suggest that $a/r = 0.61$.

Although equation (2) is semi-empirical, it is supported by the following calculations. Equation (3) may be written

$$-\frac{dm}{dt} = \frac{2\pi a m_2 K Y p_0}{k T P [K Y / (a \nu P) + 1]}, \quad (4)$$

where $K = a/r$, p_0 is the saturation vapour pressure, P the air pressure and $Y = DP$. When P is large

$$-\frac{dm}{dt} = \frac{Q = 2\pi a m_2 K Y p_0}{k T P},$$

and hence $K Y p_0$ can be calculated from the slope of $-dm/dt$ versus $1/P$. Moreover, $K Y$ can be calculated from the B term of equation (2), and hence p_0 can be calculated. This gives for 19.9°C, pot 2, di-*n*-butyl phthalate 1.42×10^{-5} mm., solid benzophenone 3.22×10^{-4} mm.; pot 4, di-*n*-butyl phthalate 1.54×10^{-5} mm., solid benzophenone 3.0×10^{-4} mm. These values are in reasonable agreement with those given in the previous section. Neumann & Völker's (1932) value for benzophenone at 19.9°C, being 3.26×10^{-4} mm. Moreover, the ratio of the diffusion coefficients for di-*n*-butyl phthalate and benzophenone is the ratio of the $K Y$ values, K being a constant, viz. $1.78/2.27 = 0.785$, which gives a reasonable value of the diffusion coefficient of benzophenone, viz. 0.049 at 1 atm., using the value for di-*n*-butyl phthalate reported later (0.0386). The agreement with the ratios of diffusion coefficients as calculated from the high-pressure slope constant A of equation (2) is, however, not satisfactory for pot 4, pot 2 giving $D(\text{di-}n\text{-butyl phthalate})/D(\text{benzophenone}) = 0.765$ and for pot 4, 0.95.

The general agreement shows that equation (3) adequately represents the rate of effusion in air, and gives extrapolated vacuum rates in agreement with experiment. It should be noticed that even when the air pressure is approximately 1000 times the vapour pressure the rate of effusion has reached approximately 20 % of the vacuum rate (cf. table 4).

THE RATE OF EVAPORATION OF DROPLETS OF DI-*n*-BUTYL PHTHALATE IN AIR

The apparatus was similar to that described in part I. Drops of radius approximately 0.45 mm. were hung from a microbalance of sensitivity 34.5 μ g./mm. deflexion. The microbalance was calibrated by means of aluminium wire and also by the buoyancy method using quartz bulbs, with good agreement.

Preliminary results showed that the original sample used in this research gave higher rates of evaporation at the higher pressures and lower rates at the lower pressures than the sample purified by molecular distillation. The $-ds/dt$ versus $1/P$ curve for the pure sample cut the curve given in part I, implying that the sample used in part I was impure. It appeared that the evaporation coefficient α was higher than the value previously published, and in order to check this point a direct determination of the vacuum rate was made.

A drop cannot be used for this purpose, owing to self-cooling in a hard vacuum. Hence a tube of radius 0.85 mm. and length 1 cm. was filled to the brim with di-*n*-butyl phthalate and the rate of evaporation in a hard vacuum was determined, using charcoal as absorbent. The liquid surface was almost flat. If A_1 = area of surface,

$$\alpha = -\left(\frac{dm}{dt}\right)_{\text{vac.}} \times \frac{1}{Ap_0} \left(\frac{2\pi kT}{m_2}\right)^{\frac{1}{2}},$$

where p_0 is the saturation vapour pressure, m_2 is the mass of evaporating molecule and m the mass of the drop. This gave $\alpha = 0.69$, i.e. greater than the value given in part I, viz. 0.28. This work was not pursued owing to the difficulty in assessing the liquid surface and errors due to wetting the glass; in the light of our more recent work the use of tricresyl phosphate rather than charcoal would have improved the method; however, the method was sufficiently accurate to show that α is near unity.

The experiments on droplets gave results of the same general type as those reported in part I. The rate of change of surface (s) with time (t) is given by

$$q = -\frac{ds}{dt} = \frac{8\pi m_2 Y c_0}{\rho P \left(\frac{Y}{P a \nu \alpha} + \frac{a}{a + \Delta} \right)}, \quad (5)$$

where $Y = DP$, ρ is the density of liquid, c_0 the saturation concentration of vapour in molecules per ml., D the diffusion coefficient at pressure P , Δ a magnitude defined in part I, $\nu = [(kT)/(2\pi m_2)]^{\frac{1}{2}}$, and a is the drop radius. A typical example is given for 15° C in table 5. Each value of ds/dt was itself obtained from approximately ten points; $q_0 = \frac{8\pi m_2 D c_0}{\rho}$.

No significant difference was obtained on repeating runs, or on replacing the absorbent by silica gel, activated alumina or Apiezon oil. This suggests that the effective pressure of di-*n*-butyl phthalate on the surface of the absorbent is zero. The results of table 5 may be represented by the equation

$$-\frac{ds}{dt} = \frac{0.920 \times 10^{-6}}{P \left(\frac{7.65 \times 10^{-4}}{aP} + \frac{1}{1 + 2.42 \times 10^{-4}/(aP)} \right)} \text{ cm.}^2 \text{ min.}^{-1} \quad (a \text{ in cm., } P \text{ in cm. of mercury})$$

TABLE 5. DIBUTYL PHTHALATE IN AIR AT 19.9°. ACTIVATED CHARCOAL ABSORBENT

P (cm. Hg)	α (mean) (cm.)	q (cm. ² min. ⁻¹) $\times 10^8$		$\frac{q_0}{q} - \frac{\alpha}{a + \Delta}$	q (calc.) (cm. ² min. ⁻¹) $\times 10^8$
		(obs.)	q_0 (cm. ² min. ⁻¹) $\times 10^8$		
8.455	0.0491	0.103	0.109	0.06	0.109
3.812	0.0486	0.293	0.241	-0.18	0.241
2.139	0.0455	0.444	0.431	-0.003	0.431
2.101	0.0481	0.398	0.438	0.10	0.438
1.163	0.0480	0.815	0.791	-0.03	0.791
0.989	0.0459	0.941	0.930	-0.01	0.930
0.656	0.0478	1.30	1.40	0.08	1.38
0.495	0.0451	1.80	1.86	0.03	1.82
0.421	0.0475	2.06	2.19	0.06	2.14
0.115	0.0470	7.20	8.00	0.157	7.30
0.0746	0.0445	11.4	12.3	0.148	10.6
0.0509	0.0461	15.3	18.1	0.276	14.7
0.0502	0.0432	14.9	18.3	0.326	14.6
0.0419	0.0438	17.3	22.0	0.386	16.9
0.0352	0.0450	19.6	26.1	0.465	19.3
0.0256	0.0437	24.5	36.0	0.643	23.9
0.0201	0.0440	29.3	45.8	0.774	27.8
0.0198	0.0449	28.7	46.5	0.833	28.3
0.0150	0.0467	32.6	61.4	1.146	33.5
0.0137	0.0422	32.6	67.2	1.342	33.2

Similar results were obtained over the temperature range 14 to 40° C. If

$$-\frac{ds}{dt} = \frac{E}{P \left(\frac{F}{aP} + \frac{1}{1 + G/(aP)} \right)} \text{ cm.}^2 \text{ min.}^{-1} \quad (a \text{ in cm., } P \text{ in cm. of mercury}). \quad (6)$$

The values of E , F and G are given in table 6. (G was calculated from the diffusion coefficient obtained from E in the manner described later. At 40° C the rates were too fast to determine F accurately.)

TABLE 6. RATES OF EVAPORATION OF DROPS OF DI-*n*-BUTYL PHTHALATE (CHARCOAL ABSORBENT)

temp. (°C)	$E \times 10^6$	$F \times 10^4$	$G \times 10^4$	diffusion coefficient at 76 cm. Hg	$S_{12} \times 10^8$ cm.
15.00	0.433	7.8	2.15	0.034	6.40
19.90	0.92	7.85	2.42	0.0386	6.32
25.00	1.86	8.15	2.57	0.0415	6.24
30.00	3.55	8.55	2.62	0.0426	6.17
35.00	6.40	8.52	2.70	0.0442	6.08
40.00	12.0	—	2.86	0.0473	6.01

The above values were calculated after a small self-cooling correction had been applied as follows. If α' is the thermal accommodation coefficient (allowing for internal degrees of freedom), it is readily shown that the heat transfer per second to the drop is

$$4\pi a^2 \alpha' P \left(\frac{2kT_1}{\pi m_1} \right)^{\frac{1}{2}} \frac{T_1 - T_0}{T_1},$$

where P is the air pressure, m_1 the mass of an 'air molecule', T_1 the temperature of the air and T_0 the temperature of the drop. If all other heat flows are neglected, this will be balanced by the rate of loss of heat by evaporation, once the drop has attained the steady state, i.e.

$$-\frac{L}{M} \frac{dm}{dt} = 4\pi a^2 \alpha' P \left(\frac{2kT_1}{\pi m_1} \right)^{\frac{1}{2}} \frac{T_1 - T_0}{T_1}$$

or
$$\delta T = T_1 - T_0 = \frac{\rho L}{8aP\alpha'} \left(\frac{m_1 T_1}{2\pi k} \right)^{\frac{1}{2}} q,$$

where $q = -ds/dt = -\frac{2}{a\rho} \frac{dm}{dt}$. Taking $\alpha' = 1$ for a liquid surface

$$\delta T = 2.18q/(aP) \text{ at } 25.00^\circ \text{C} \quad \text{and} \quad \delta T = 2.19q/(aP) \text{ at } 30.00^\circ \text{C}.$$

The largest corrections are of the order of 5 % decrease in observed evaporation rate (for the lowest pressures), and scarcely affect the results.

From the values of E and the vapour pressures the diffusion coefficients may be calculated, since $E = \frac{8\pi m_2 D c_0}{\rho}$. These are listed in table 6; the value of D at 19.9°C is slightly greater than that reported in part I. It will be observed that D increases as T increases. This implies that the collision radius decreases as T increases owing to the increased kinetic energy of impact. The collision radius S_{12} , i.e. the sum of the radii of gas molecule and di-*n*-butyl phthalate, has been calculated from the equation

$$D = \frac{2}{3} \frac{1}{(1 + \alpha_{21})} \frac{1}{\pi n S_{12}^2} \left[\frac{2kT(m_1 + m_2)}{\pi m_1 m_2} \right]^{\frac{1}{2}},$$

where n is the number of molecules per cm^3 , m_1 and m_2 the masses of 'air molecules' and evaporating molecules (cf. Jeans 1940), taking $\alpha_{21} = \frac{1}{3}$, and is listed in table 6. When $\log D$ is plotted against $\log T$ a straight line is obtained, from the slope of which it appears that D is proportional to T^3 , likewise $S_{12} \propto T^{-\frac{1}{2}}$. This temperature dependence for D is rather greater than would be expected theoretically, simple kinetic theory giving $D \propto T^{\frac{1}{2}}$ at constant pressure, while for molecules repelling according to the inverse β power of the distance, $D \propto T^{\left(\frac{3}{2} + \frac{2}{\beta} - 1\right)}$ (cf. Chapman & Cowling 1939, p. 248). Sutherland's formula gives $D = D_0 T^{\frac{1}{2}}/(1 + C/T)$ at constant pressure, where C is a constant, or $D = D_0 T^{\frac{1}{2}}/(T + C)$. Thus the maximum power of T in the empirical expression $D = D_0 T^x$ is $\frac{5}{2}$, obtaining when $C \gg T$, and the minimum is $\frac{3}{2}$ (for $C \ll T$). When the force between two molecules contains an attractive and repulsive term, and is of the form $\lambda/r^\beta - \mu/r^\gamma$, where the repulsive exponent β is $> \gamma$, the attractive exponent, and λ and μ are constants, Sutherland's equation becomes

$$D = \frac{D_0 T^{\left(\frac{3}{2} + \frac{2}{\beta-1}\right)}}{\left(1 + \frac{C}{T^{\beta-1}}\right)}$$

(cf. Chapman & Cowling, p. 181), and hence, if $\gamma = 7$,

$$D = \frac{D_0 T^{\left(\frac{3}{2} + \frac{2}{\beta-1} + \frac{\beta-7}{\beta-1}\right)}}{\left(C + T^{\frac{\beta-7}{\beta-1}}\right)}.$$

It is clear that the maximum value of x in the expression $D = D_0 T^x$ will occur when $T^{\frac{\beta-7}{\beta-1}} \ll C$ and $\beta \gg 7$, and will be $\frac{5}{2}$. Hence allowance for attractive forces will not account for the observed value, $x = 3$.

While it must be remembered that the calculation of diffusion coefficient involves a knowledge of the vapour pressure, and that only small changes occur in D over the temperature range considered, it must also be borne in mind that most work on D has been done on small molecules, and theoretical work is confined to simple molecules. A polyatomic molecule such as di-*n*-butyl phthalate might well be more deformable and D might well be more temperature-sensitive than, say, an argon atom.

The values of S_{12} deduced above are reasonable, and on subtraction of 1.87 Å for air (Jeans 1940) give reasonable values for the size of a di-*n*-butyl phthalate molecule, e.g. at 19.90°, radius = 4.45 Å. The value deduced from the density of the liquid, assuming spherical close-packed molecules, is 4.3 Å, and agrees with the value used in part I, deduced from the dimensions 6.28 – 1.87 = 4.41 Å.

This value of S_{12} deduced from D was used in the calculation of $\Delta = kT/(\pi S_{12}^2 P)$, i.e. Δ was obtained from the constant E . It should be noted that even if $\Delta = 0$ the ds/dt versus $1/P$ curve passes asymptotically to the vacuum rate, when an equation of the same form as equation (2) is obtained. It is found, however, that the use of an equation in which $\Delta = 0$ leads to values of $\alpha = 1.29$, i.e. greater than unity, which is impossible; to this extent the theory which introduces Δ is justified.

In addition to the diffusion coefficient the evaporation coefficient may be calculated. If q is the observed value of $-ds/dt$ and q_0 the value which would be obtained if Langmuir's theory applied (cf. part I), then $\frac{q_0}{q} - \frac{a}{a+\Delta}$ is plotted against $1/P$, and the slope

$$d\left(\frac{q_0}{q} - \frac{a}{a+\Delta}\right) / d\left(\frac{1}{P}\right) = \frac{Y}{a\nu\alpha},$$

whence α is readily calculated, since $Y = DP$. Results are given in table 7.

TABLE 7. EVAPORATION COEFFICIENT

temp. (°C)	slope = $Y/(a\nu\alpha)$	a (cm.)	$Y = DP$	ν	α	α smoothed
15.00	0.018	0.0434	2.59	3702	0.90	0.97
19.90	0.017	0.0450	2.93	3732	1.03	1.00
25.00	0.0187	0.0436	3.15	3763	1.03	1.00
30.00	0.0202	0.0423	3.24	3795	1.00	1.00
35.00	0.0198	0.0430	3.36	3826	1.03	1.05

The smoothed values of α were calculated from the smoothed values of D obtained by plotting $\log D$ versus $\log T$. It is clear that α may be taken to be unity over the whole temperature range.

In addition, a further check on the latent heat of evaporation may be calculated. The vacuum rates of evaporation may be calculated from the equation for $-ds/dt$ (equation (6)), viz. $-ds/dt = Ea/F$ at hard vacuum. Hence per unit area

$$-\frac{dm}{dt} = \frac{Ea}{F} \frac{\alpha \rho}{2} \frac{1}{4\pi a^2} = \frac{E\rho}{8\pi F}$$

for a hard vacuum. Theoretically this value $= \frac{m_2 p_0 \alpha}{(2\pi m_2 kT)^{1/2}}$, and hence

$$\frac{d \log p_0}{d(1/T)} = -\frac{L}{R} = \frac{d}{d(1/T)} \left[\log \left(-\frac{dm}{dt} \right) + \frac{1}{2} \log T \right],$$

taking $\alpha = 1$. Hence L , the latent heat per mole, may be calculated from the slope of $\log (-dm/dt) + \frac{1}{2} \log T$ plotted against $1/T$. This gives $L = 21,900$ cal. per mole in exact agreement with the value deduced from the vapour-pressure measurements. In addition, L may be calculated from the Langmuir portion of the $(-ds/dt)$ versus $1/P$ curve, since in this region $-\frac{ds}{dt} = \frac{8\pi MDp_0}{RT\rho}$, and since $D \propto T^3$, the plot of $\log (-ds/dt) - 2 \log T + \log \rho$ versus $1/T$ enables L to be calculated, the value 21,900 being again obtained.

THE RATE OF EVAPORATION OF DROPLETS OF DI-*n*-BUTYL PHTHALATE IN HYDROGEN AND FREON

Preliminary work with hydrogen using charcoal absorbent showed that the expected increase in diffusion coefficient occurred, but that the evaporation coefficient was reduced to 0.6. Since the latter should be independent of the ambient gas, either the theory was incorrect or the absorbent was at fault. The use of silica gel and activated alumina as absorbents gave the expected result, $\alpha = 0.99$ and 1.01 respectively at 19.9° C, whereas Apiezon oil on glass-wool gave a slightly lower value of α (0.9), but the same evaporation coefficients at the high pressure. Presumably the charcoal desorbs air slowly in spite of the technique used of repeated fillings with hydrogen and pumping down.

With Freon 12 the same technique was employed, but it was found that Apiezon oil was the best absorbent. With silica gel the Langmuir region evaporation characteristics were the same as for Apiezon oil, but the low-pressure region (0.05 to 0.01 cm.) gave lower rates, giving $\alpha = 0.7$ at 19.90° C. The absorption of the readily condensible Freon on the silica gel was, moreover, too great to allow steady pressure readings to be taken, giving a wide scatter to the points. The results using the best absorbents are summarized in table 8, which gives the constants E , F and G of equation (6).

The values of D , α and S_{12} were calculated as in the previous section. The collision radius of H_2 may be taken as 1.36 Å (Jeans), and the radius of Freon estimated from the electron diffraction data of Brockway (1937) is 3.2 Å. The value computed from the density for close-packed spheres is 3.0 Å. Subtracting these values (3.2 Å for Freon) from the values of S_{12} given in table 8 we obtain the following values for the radius of the di-*n*-butyl phthalate molecule at 19.90° C: from the measurements

TABLE 8. EVAPORATION OF DROPS OF DI-*n*-BUTYL PHTHALATE
IN HYDROGEN AND FREON

temp. (°C)	gas	α (mean) in cm.	$E \times 10^6$	$F \times 10^4$	$G \times 10^4$	$F \times 10^4$ smoothed	D at 76 cm.	α	$S_{12} \times 10^8$ cm.
19.90	H ₂	0.0443	3.65	31.2	2.61	31.2	0.153	0.99	6.04
19.90	Freon	0.0448	0.30	2.57	1.36	2.57	0.0126	1.00	8.44
30.00	Freon	0.0428	1.14	2.98	1.48	2.80	0.0140	0.95	8.21

using H₂ 4.68 Å, from the measurements using Freon 5.24 Å; the difference is possibly due to the size of the Freon molecule, which does not penetrate the di-*n*-butyl phthalate molecule as deeply as 'molecules of air'. The agreement for hydrogen and air supports the use of the factor $\left(\frac{m_1 + m_2}{m_1 m_2}\right)^{\frac{1}{2}}$ which appears in all formulae for the diffusion coefficient, although the formula for rough spheres contains a further term dependent on m_1 and m_2 involving radii of gyration (cf. Chapman & Cowling). The agreement in the value of α for hydrogen, air and Freon is good support for the theory.

THE SHAPE OF THE DROPS

The drops used were slightly smaller than those described in part I, and some doubt was felt as to their shape; up to a certain size the *larger* the drop the more spherical is the shape, when the drop is hanging at the top of a quartz fibre. Drops of radii 0.55, 0.48 and 0.42 mm. were suspended from the top of the fibre used for evaporation, and shadow photographs were taken using an enlarger (magn. $\times 16$), with a glass scale for calibration. The mean radius was calculated from the mass, and hence the surface, assuming a spherical drop. This was compared with the surface computed from the dimensions, assuming an ellipsoid of revolution. It was found that the errors involved in assuming spherical shape were less than 1 %.

DISCUSSION

The value of α recorded in this paper for di-*n*-butyl phthalate is unity, but the work of Alty (1935, 1937) and Baranaev (1946) shows that other polar molecules such as water and the aliphatic alcohols have much lower values of α . It should be noted, however, that a liquid such as di-*n*-butyl phthalate is ideal for the study of evaporation coefficients, since rates of evaporation are low enough for self-cooling to be negligible under proper conditions, and since the vapour pressure is so low that rates of evaporation at very low air pressures are not influenced by the presence of vapour molecules. The last point is of some importance with more volatile liquids. Our formulae, derived in part I, are readily applicable to a plane or approximately plane surface of area A_1 , for which the vapour diffuses away from a region of concentration c_1 molecules per ml. given by

$$A_1 \nu \alpha (c_0 - c_1) = D \frac{A_1 c_1}{(l - \Delta)},$$

where c_0 is the saturation concentration in molecules per ml., D the diffusion coefficient at the air pressure used, l the length of the diffusion gradient, i.e. the distance from the liquid surface to the absorbing surface. Solving for c_1 as before, this gives for the rate of evaporation

$$\frac{DA_1}{(l-\Delta)} \frac{c_0 \nu \alpha m_2}{\nu \alpha + \frac{D}{(l-\Delta)}} = \frac{A_1 c_0 \nu \alpha m_2}{1 + \frac{D}{\nu \alpha (l-\Delta)}} \text{ g./sec.}$$

A similar formula has been derived by Brookfield, Fitzpatrick, Jackson, Matthews & Moelwyn-Hughes (1947), who used liquids in tubes, and tested the Langmuir region by absorbing the vapour in charcoal, or in sulphuric acid (in the case of water vapour) and determining the rate of evaporation as a function of l in air at atmospheric pressure. However, only the Langmuir region was studied. A similar experimental method was used by Langmuir & Schaeffer (1943). When the term $\frac{\nu \alpha (l-\Delta)}{D}$

is large compared with unity, the rate of evaporation will be given by $\frac{Ac_0 D m_2}{(l-\Delta)}$, i.e. the 'Langmuir rate'; l would have to be very small in order that $\frac{\nu \alpha (l-\Delta)}{D}$ should be unity, since ν is approx. 10^4 for water, D is approx. 0.25 in air at 1 atm., and even if α is 0.04 (Alty 1937), $(l-\Delta)$ would have to be 6×10^{-4} cm. before the rate of evaporation began to approach the vacuum rate. Thus very small values of l would have to be used in order to determine α by this method; Brookfield *et al.* used 1.6 cm. as the smallest value of l . It might be possible to use the method at lower air pressures, e.g. at 7.6 mm. of air $(l-\Delta) = 0.6$ mm. for $\frac{\nu \alpha (l-\Delta)}{D} = 1$.

The low values of α , if real, imply some impedance to the condensation of vapour, which is probably most simply thought of as (a) an energy barrier and (b) an orientation factor such that favourable orientation is necessary for condensation. The combination of (a) and (b) would give α of the form $\alpha = \theta e^{-E'/RT}$, where θ is the orientation factor, which may alternatively be thought of as an entropy term, and E' the activation energy necessary to pass the barrier. The existence of an E' term is perhaps most clear in the work of Langmuir on the evaporation of water through monolayers spread on the water; Langmuir & Schaeffer (1943), however, did not calculate the value of the θ factor. It should be noted that the vapour-pressure equation will not involve E' , since the equations for evaporation and condensation will each involve a term $e^{-E'/RT}$. It is hoped in the future to make a theoretical study of the value of α for complex molecules by the transition state method.

The vacuum rate of loss of mass in g./cm.²/sec. during evaporation may be calculated from the data above to be $2 \times 10^{10} e^{-21,900/RT}$, and the vapour pressure in mm. of mercury to be $3.5 \times 10^{11} e^{-21,900/RT}$. The entropy of vaporization ΔS_p , taking unit pressure as 1 mm. of mercury, is given by $T\Delta S_p = \Delta H + RT \log_e p_{mm}$ and from our data at 25.00° C $\Delta S_p = 0.0527$ (kcal./degree), in reasonable agreement with the value deduced from the Barclay-Butler (1938) rule, $\Delta S_p = 0.0277 + 0.0011L$, which gives $\Delta S_p = 0.0518$.

This paper is based on work which was submitted as a thesis by one of us (J.B.) for partial requirement for the Ph.D. degree at the University of Leeds. Our thanks are due to Professor M. G. Evans, F.R.S., for his constant interest and encouragement.

REFERENCES

- Alty, T. 1937 *Proc. Roy. Soc. A*, **161**, 68.
Alty, T. & Mackay, C. A. 1935 *Proc. Roy. Soc. A*, **149**, 104.
Baranaev, M. K. 1946 *J. Phys. Chem. Soc.* **20**, 399.
Barclay, I. M. & Butler, J. A. V. 1938 *Trans. Faraday Soc.* **34**, 1445.
Bradley, R. S., Evans, M. G. & Whytlaw-Gray, R. W. 1946 *Proc. Roy. Soc. A*, **186**, 368.
Brockway, L. O. 1937 *J. Phys. Chem.* **41**, 185, 746.
Brookfield, K. J., Fitzpatrick, A. D. N., Jackson, J. F., Matthews, J. B. & Moelwyn-Hughes, E. A. 1947 *Proc. Roy. Soc. A*, **190**, 59.
Burrows, G. 1946 *J. Soc. Chem. Ind., Lond.*, **65**, 360.
Chapman, S. & Cowling, T. G. 1939 *The mathematical theory of non-uniform gases*. Cambridge University Press.
Hickman, K. C. D. 1936 *J. Franklin Inst.* **221**, 383.
Hickman, K. C. D., Hecker, J. C. & Embree, N. D. 1937 *Industr. Engng Chem. (Anal. ed.)*, **9**, 264.
Jeans, J. 1940 *An introduction to the kinetic theory of gases*. Cambridge University Press.
Kapff, S. F. & Jacobs, R. B. 1947 *Rev. Sci. Instrum.* **18**, 581.
Langmuir, J. & Schaeffer, V. J. 1943 *J. Franklin Inst.* **235**, 119.
Neumann, K. & Völker, E. 1932 *Z. Phys. Chem. A*, **161**, 33.
Verhoek, F. H. & Marshall, A. L. 1939 *J. Amer. Chem. Soc.* **61**, 2737.
Woodland, D. J. & Mack, E. 1933 *J. Amer. Chem. Soc.* **55**, 3149.
Zabel, R. M. 1933 *Rev. Sci. Instrum.* **4**, 233.

The rate of evaporation of droplets. III. Vapour pressures and rates of evaporation of straight-chain paraffin hydrocarbons

BY R. S. BRADLEY AND A. D. SHELLARD

Department of Inorganic and Physical Chemistry, University of Leeds

(Communicated by M. G. Evans, F.R.S.—Received 28 December 1948)

The vapour pressures, latent heats of vaporization and fusion and diffusion coefficients of the vapours in air, have been determined at 15 to 40° C, by a combination of Knudsen's vapour-pressure technique and studies on the rate of evaporation of drops and solid beads, for $n\text{-C}_{18}\text{H}_{38}$, $n\text{-C}_{17}\text{H}_{36}$ and $n\text{-C}_{16}\text{H}_{34}$. The rates of evaporation of drops and solid beads agree with the previously published theory, and lead to a value of unity for the evaporation coefficient. A new experimental method is described for determining the self-cooling on evaporation of a drop. The bearing of the results on the size and shape of hydrocarbon molecules is discussed.

INTRODUCTION

In part I (Bradley, Evans & Whytlaw-Gray 1946) and part II (Birks & Bradley 1949) a description has been given of experiments on the vapour pressure and evaporation rate of di-*n*-butyl phthalate. In this work the application of the theory of Fuchs

(1934) to droplet evaporation was established. The present paper gives additional support to the theory, but the main purpose of the work was to apply the technique to a number of related molecules. The higher straight-chain paraffin hydrocarbons were chosen for this comparative study, not only on account of the interest attaching to the molecular configuration in the liquid and gaseous state, but also because of gaps in the published data for these substances. The values of the collision radii we have obtained and the mean configuration we consider probable are also of interest in kinetic studies. It is hoped later to extend the work to branched-chain paraffin hydrocarbons and to fluorinated hydrocarbons and hydrocarbon mixtures.

PREPARATION AND PHYSICAL PROPERTIES OF THE HYDROCARBONS

n-Hexadecane was prepared from cetyl iodide by reduction, and also by the Wurtz synthesis from *n*-octyl iodide; *n*-octadecane was prepared from *n*-octadecyl alcohol, via the iodide. The Carbide and Carbon Chemicals Corporation of the U.S.A. very kindly gave us a sample of pure *n*-heptadecane, and we are also grateful to Professor Ubbelohde for a small sample of pure octadecane which was useful for comparison purposes. The final purification of the samples we prepared was made by molecular distillation.

The following physical constants were determined: refractive index n_D^{20} : $C_{16}H_{34}$ 1.43469 (from the Wurtz reaction, which gave the purer product), $C_{17}H_{36}$ n_D^{25} 1.4348 (determined by Dr Audsley of the Department of Organic Chemistry), $C_{18}H_{38}$ n_D^{30} 1.44295; density $C_{16}H_{34}$ ρ_4^{20} 0.7743 $C_{17}H_{36}$ ρ_4^{25} 0.7745, ρ_4^{30} 0.7711, ρ_4^{35} 0.7676 (determined by Dr Audsley), $C_{18}H_{38}$ ρ_4^{35} 0.7721. These densities were necessary for the calculation of the drop surface, and since some experiments were made on solid hydrocarbon beads the density of the solids was also determined. Some difficulty was experienced in the use of a density bottle containing water, owing to air which was trapped by the greasy surface and which was difficult to remove even on pumping, and we therefore used a method we have not seen described in the literature, viz. the method of air buoyancy. Approximately 0.1 g. of hydrocarbon was suspended from the quartz microbalance of the type described in part I, and was counterbalanced by means of a quartz weight suspended from the other arm of the balance. The change in deflexion on changing the pressure was followed on a travelling microscope, and from the density of the quartz counterpoise that of the solid hydrocarbon could readily be calculated. With 0.15 g. of hydrocarbon a change of 1 atm. gave a deflexion of the order of 2 mm., which could be read to 0.002 mm. on the travelling microscope. Results are as follows: $C_{16}H_{34}$ ρ_4^{15} 0.865 ± 0.001 , $C_{17}H_{36}$ ρ_4^{15} 0.864 ± 0.004 , $C_{18}H_{38}$ ρ_4^{15-20} 0.873 ± 0.004 .

As a further check on purity X-ray powder photographs of the solids were taken by Mr S. C. Nyburg, by kind permission of Professor E. G. Cox. The specimens were melted and introduced into a cellophane tube of diameter 1.5 mm., mounted in a camera of 9.6 cm. radius. Unfiltered copper radiation was used for an exposure of 1 hr. As far as could be detected the sample of octadecane supplied by Professor Ubbelohde was identical with ours; the same remark applies to the evaporation characteristics.

VAPOUR PRESSURES AND HEATS OF VAPORIZATION

The vapour pressures were determined by the Knudsen method, as described in parts I and II, using benzophenone as the calibrating substance. It was necessary, in the determination of the vapour pressure of the solid hydrocarbons, to allow them to solidify under vacuum to prevent explosive degassing of the solid. The effusion tubes were mounted in a solid aluminium block, and self-cooling appeared to be absent, since the same results for vapour pressure, within the experimental error, were obtained with different-sized holes. This independence of hole size also suggests that the correct conditions for mean free path were maintained. According to Knudsen (1909) the mean free path of the effusing molecules should be at least ten times the hole diameter, from his observations on the effusion of carbon dioxide from a higher to a lower pressure. It is not clear, however, how Knudsen's experimental result applies to our mode of operation, since the vapour in our experiments was maintained at constant pressure in the tube and effused into a hard vacuum. As a criterion we have calculated from the evaporation data (*vide infra*) the mean free paths of the hydrocarbon molecules for collision with like molecules, viz. $\lambda_{11} = \frac{1}{\sqrt{2}} \frac{kT}{\pi p_0 S_{11}^2}$,

where p_0 is the vapour pressure and S_{11} the diameter of a hydrocarbon molecule, and in every instance λ_{11} is greater than the hole diameters (0.14, 0.10 and 0.06 cm.), although not always ten times. Recently Zilberman-Granovskaja & Schugan (1940) have found, with a technique similar to ours, that ideal effusion will occur provided that the mean free path is greater than one-fifth the diameter of the hole, a much less severe condition than Knudsen's. It should be noticed, however, that departure from the true result with holes which are too large may occur through self-cooling or from a departure from ideal molecular flow, and it is not easy to disentangle the two factors. It should be noted, moreover, that the results are less dependent on the hardness of the vacuum than may be supposed, since we have found that an external pressure of 1000 times the vapour pressure reduces the rate of effusion by only 80 % (part II). Some workers have subtracted the small air pressure in the apparatus from the vapour pressure, in order to correct for the imperfection of the vacuum, but this is clearly incorrect; this again emphasizes the difference between our technique and that due to Knudsen.

Graphs of $\log p_0$ when plotted against $1/T$ gave good straight lines, in further evidence of the correctness of the conditions of flow. The constants A and B in the equation $\log_{10} p_0 = \frac{-A}{T} + B$, p_0 in cm. of mercury, and the latent heats of fusion, L_F , of vaporization of the solid, L_S , and of vaporization of the liquid, L , in calories per mole, are given in table 1.

For $n\text{-C}_{16}\text{H}_{34}$ and $n\text{-C}_{17}\text{H}_{36}$ the latent heat for the solid has been calculated from two points only, whereas six points at different temperatures could be plotted for most of the other cases. The published values for the vapour pressures at room temperature are very scanty. Ubbelohde (1938) studied n -hexadecane, melting-point 18.1°C , at 9 and 15°C , and obtained consistent results at 9°C with different hole sizes, but at 15°C the use of holes of radii 0.33, 0.58 and 0.75 mm. gave vapour

pressures of 1.38, 1.04 and 0.56 dynes cm.⁻² respectively. Ubbelohde believed that the trend of the results at 15° C was related to premelting. To test this point we studied *n*-hexadecane at 16.97° C, i.e. very near to the melting-point, and observed no abnormal behaviour in the vapour pressure and rate of evaporation; so far as our measurements go there is no evidence of premelting. Ubbelohde's value for solid *n*-hexadecane at 9° C may be extrapolated to 14.97° C, using his value of the latent heat of vaporization of the solid, 28,800 cal./mole; to give a vapour pressure of 0.508×10^{-4} cm. of mercury, in reasonable agreement with our value, 0.43×10^{-4} cm. of mercury, considering the nature of the extrapolation.

TABLE 1. VAPOUR PRESSURE CONSTANTS AND LATENT HEATS

hydro- carbon	state	range of temp. (° C)	A	B	L	L_g	L_F
<i>n</i> -C ₁₆ H ₃₄	liquid	20-35	4189	10.260	19,170 ± 600	—	—
	solid	15-17	6579	18.466	—	30,100 ± 1000	10,930 ± 1600
<i>n</i> -C ₁₇ H ₃₆	liquid	25-40	4374	10.333	20,020 ± 700	—	—
	solid	15-20	6866	18.738	—	31,420 ± 1100	11,400 ± 1800
<i>n</i> -C ₁₈ H ₃₈	liquid	30-40	4730	11.036	21,640 ± 800	—	—
	solid	15-25	7995	21.831	—	36,580 ± 1200	14,940 ± 2000

Our values of the latent heats of vaporization of the liquid hydrocarbons are in good agreement with those published by Rossini (1945), viz. C₁₆H₃₄ 19,360, C₁₇H₃₆ 20,540, C₁₈H₃₈ 21,720 cal./mole. These and other values of Rossini support the view that the latent heat of vaporization of the liquids is a linear function of the number (*n*) of carbon atoms in the molecule. Huggins (1939), however, claimed that unpublished data of Rossini supported his view that on vaporization the hydrocarbon molecule became coiled into a ball, and that the latent heat of vaporization of the liquid was proportional to $n^{\frac{2}{3}}$; the more recent data of Rossini do not support this relation, but leave open the possibility of coiling.

THE RATE OF EVAPORATION OF DROPS AND BEADS OF HYDROCARBONS

The rate of evaporation of drops on to absorbents of charcoal and silica gel was studied as a function of the air pressure over a range of temperature in the manner described in parts I and II. As explained in these earlier papers the rate of change of mass with time is given by

$$-\frac{dm}{dt} = \frac{4\pi a D c_0 m_2}{\frac{D}{a\nu\alpha} + \frac{a}{a+\Delta}},$$

and the rate of change of surface with time by

$$-\frac{ds}{dt} = \frac{8\pi Dc_0m_2}{\rho \left[\frac{D}{a\nu\alpha} + \frac{a}{a+\Delta} \right]},$$

where m is the mass, s the surface, a the radius and ρ the density of the drop, D the diffusion coefficient at the air pressure used, c_0 the saturation concentration of hydrocarbon vapour in molecules per ml., m_2 the mass of the hydrocarbon molecule, α the evaporation coefficient, $\nu = \left(\frac{kT}{2\pi m_2} \right)^{\frac{1}{2}}$, k the Boltzmann constant, and Δ the effective thickness of the vacuous envelope surrounding the drop.

The study of droplet evaporation thus falls into two sections, the first when $D/(a\nu\alpha)$ is small compared with $a/(a+\Delta)$, which is then unity, when the 'Langmuir' rate is observed, and the diffusion coefficient may be calculated; the second when $D/(a\nu\alpha)$ cannot be neglected in comparison with $a/(a+\Delta)$, and the rate of evaporation approaches the vacuum rate. The second region is more difficult to study than with di-*n*-butyl phthalate (cf. parts I and II) owing to the greater vapour pressure of the hydrocarbons, and the greater rate of evaporation at the lower pressures; silica gel was found to be the best absorbent in this region. It is therefore not possible to study an isolated hydrocarbon drop at the lowest air pressures without making some allowance for self-cooling, a special study of which was therefore made.

The diffusion coefficient may readily be calculated from the rate of evaporation in the 'Langmuir' region, when $-ds/dt = q = 8\pi Dc_0m_2/\rho = \text{constant}/P$. The values of qP are given in table 2.

TABLE 2. qP IN 10^{-4} CM.² MIN.⁻¹ \times CM. OF MERCURY

temp. (°C)	<i>n</i> -C ₁₆ H ₃₄	<i>n</i> -C ₁₇ H ₃₆	<i>n</i> -C ₁₈ H ₃₈
14.97	0.278	0.058	0.00889
16.97	0.395	—	—
20.01	0.671	0.153	0.02667
24.98	1.168	0.3658	0.07900
29.98	2.095	0.6593	0.217
34.88	3.535	1.152	0.3970
39.94	—	1.967	0.725

The melting-points of *n*-C₁₆H₃₄, *n*-C₁₇H₃₆ and *n*-C₁₈H₃₈ are respectively 18.1, 22.0 and 28.1° C, so that some of the above data refers to solid hydrocarbons.

From the vapour pressures listed above, the diffusion coefficient in 76 cm. of air may be calculated. When $\log D$ is plotted against $\log T$ a fair straight line is obtained, from which it appears that $D \propto T^2$, i.e. a smaller temperature dependence than for di-*n*-butyl phthalate (cf. part II). The lines for liquid and solid hydrocarbons show a slight separation, which we attribute to surface irregularities on the solid bead. The separation suggests that the rates of evaporation from the solid should be reduced by 2 or 3 % to obtain the values for a perfectly smooth bead. The smoothed values for D listed in table 3 are taken from the plot of $\log D$ versus $\log T$, and allow for surface irregularity.

TABLE 3. DIFFUSION COEFFICIENT AT 76 CM. OF AIR, IN C.G.S. UNITS

temp. (°C)	$n\text{-C}_{16}\text{H}_{34}$		$n\text{-C}_{17}\text{H}_{36}$		$n\text{-C}_{18}\text{H}_{38}$	
	exp.	smoothed	exp.	smoothed	exp.	smoothed
14.97	0.0384	0.0380	0.0403	0.0395	0.0388	0.0373
16.97	0.0392	0.0385	—	—	—	—
20.01	0.0392	0.0393	0.0422	0.0408	0.0396	0.0385
24.98	0.0402	0.0405	0.0423	0.0423	0.0418	0.0397
29.98	0.0425	0.0520	0.0436	0.0438	0.0408	0.0411
34.88	0.0435	0.0435	0.0453	0.0452	0.0421	0.0425
39.94	—	—	0.0463	0.0466	0.0445	0.0439

It will be seen that the values of D for the three hydrocarbons at the same temperature are very near to one another, the extreme differences being only 6 %. $\text{C}_{18}\text{H}_{38}$ and $\text{C}_{16}\text{H}_{34}$ lie in the expected relation to one another, but $\text{C}_{17}\text{H}_{36}$ appears out of order. In view of the nature of the calculation of D from two experimental quantities, the rate of evaporation and the vapour pressure, both of which may be in error to a few per cent, it is not possible to state definitely whether or not the position of $\text{C}_{17}\text{H}_{36}$ is real and related to the odd number of carbon atoms. The significance of the approximate equality of the value of D for the three hydrocarbons is discussed in the last section.

In contrast to the requirements for the diffusion coefficient, evaluation of the evaporation coefficient requires measurements to be made at comparatively low pressures of the order of 0.1 mm. of mercury, when the hydrocarbons studied evaporate sufficiently fast for the self-cooling to be an important factor (contrast di-*n*-butyl phthalate, part II). This difficulty, for the liquid, is not removed by studying hydrocarbons of higher molecular weight, since the melting-point is raised and higher temperatures must be employed. The self-cooling, however, will be less important for suitable branched-chain hydrocarbons, which have a lower melting-point than their straight-chain isomers, and which, it is hoped, will be studied in the future. In view of the uncertainties attaching to the thermal accommodation coefficient, and the contributions to thermal flow of conduction through the fibre and of radiation, an experimental study was made of self-cooling.

A droplet of octadecane of the same size as that used for evaporation studies was suspended from a short fibre in the balance case, above a watch-glass containing silica gel, the micro-balance having been removed, and was observed by means of a microscope. The essence of the method consisted in maintaining the temperature of the thermostat at a value slightly above the melting-point, and in reducing the pressure inside the case until, with a suitable adjustment of pressure and temperature, self-cooling just caused solidification; the self-cooling is then the difference in temperature between the melting-point and the temperature of the thermostat. In operation the method was complicated by the appearance of a second crystalline form of octadecane. The melting-point of the stable form is 28.1° C, but the drop did not solidify when kept for several hours at 28.0° C at atmospheric pressure, or for several hours at intermediate temperatures between 28.1 and 26.8 to 26.9° C. Solidification to a transparent solid occurred, however, on cooling below 26.8 to 26.9° C, and this solid reverted to the normal opaque form on rapid cooling. On

heating the transparent solid it melted at 26.8 to 26.9° C, whereas the normal opaque form always melted at 28.1° C. This behaviour was strictly reproducible, and it was concluded that a metastable form of melting-point 26.8 to 26.9° C was formed by slow cooling of the liquid. In confirmation of this view Carey & Smith (1933) prepared a sample of octadecane of melting-point 27.9 to 28.0° C, and a second form with a melting-point at *c.* 27.1° C, and Ubbelohde (1938) showed that an arrest occurred in the cooling curve at 27.0° C, and no viscous flow below 26.85° C.

Because of this transition it was necessary to cool the liquid to just above the *metastable* melting-point 26.8 to 26.9° C. The pressure was then reduced, kept constant at a low value, and the drop was observed for signs of solidification. If no solidification occurred after 30 min. the pressure was reduced, kept constant, and the drop was again observed. If no solidification occurred at the lowest pressure used in the evaporation studies, the whole process was repeated at a series of lower temperatures, until solidification occurred. Values of the radius of the drop were also recorded at intervals. Results are given in table 4.

TABLE 4. SELF-COOLING OF DROPS

temp. (°C) (thermostat)	pressure in cm. Hg	nature of drop	radius of drop (cm.)	self-cooling (°C)
27.7	<0.001	liquid	—	—
27.5	<0.001	liquid	—	—
27.4	0.0089	solid	0.050	0.5-0.6
27.3	0.01015	solid	0.057	0.4-0.5

The maximum theoretical self-cooling was calculated in the manner described in part II, on the assumption that the thermal accommodation coefficient is unity, no allowance being made for conduction through the fibre or for radiation transfer of heat, and was found to be approximately five times the above experimental values. This factor was assumed to be applicable to all conditions, and the rates of evaporation corrected accordingly. In addition, for comparison, the rates were corrected on the assumption of 'theoretical' self-cooling.

As in part II, the results for the evaporation of drops could be represented by the equation

$$-\frac{ds}{t} = \frac{E}{P \left[\frac{F}{\alpha P} + \frac{1}{1 + G/(\alpha P)} \right]} \text{ cm}^2 \text{ min.}^{-1} \quad (\alpha \text{ in cm., } P \text{ in cm. of mercury}),$$

where *E*, *F* and *G*, corrected for 'theoretical' and 'experimental' self-cooling, are constants given in table 5, together with the corresponding values of the evaporation coefficient α , calculated as in part II. The greatest correction due to self-cooling occurred with *n*-octadecane at the lowest pressure, 0.0478 cm. of mercury, when $-ds/dt$ observed was $34.95 \times 10^{-5} \text{ cm}^2 \text{ min.}^{-1}$, corrected for 'theoretical' self-cooling $35.98 \times 10^{-5} \text{ cm}^2 \text{ min.}^{-1}$, corrected for 'experimental' self-cooling $35.17 \times 10^{-5} \text{ cm}^2 \text{ min.}^{-1}$. At higher pressures the corrections were correspondingly smaller. It will be seen that the 'experimental' self-cooling brings the evaporation coefficient to unity or less; the 'theoretical' self-cooling, however,

gives values slightly greater than unity, which is an impossible result, and to this extent our procedure is justified. Although it was not possible, because of the greater rates of evaporation, to study octadecane and heptadecane over the same range of pressure as di-*n*-butyl phthalate, our results show that the evaporation coefficient may be taken as unity both for the liquid and solid straight-chain hydrocarbons.

TABLE 5. CONSTANTS IN THE EVAPORATION EQUATION

hydro-carbon	temp. (°C)	$E \times 10^6$	$F \times 10^4$		$G \times 10^4$	α	
			exp. self-cooling	theoretical		exp. self-cooling	theoretical
<i>n</i> -C ₁₈ H ₃₈	14.97	0.889	7.35	7.00	2.295	1.00	1.05
	20.01	2.667	7.60	6.80	2.342	0.99	1.10
	29.98	21.7	8.62	7.75	2.460	0.91	1.02
<i>n</i> -C ₁₇ H ₃₆	14.97	5.80	7.85	7.05	2.413	0.96	1.07

DISCUSSION ON THE SIZE AND SHAPE OF HYDROCARBON MOLECULES

As described in part II, values of S_{12} the 'collision radius', i.e. the sum of the radii of hydrocarbon and 'air molecules', were obtained from the relation

$$S_{12}^2 = \frac{2}{3\pi D_{12}} \frac{1}{1 + \alpha_{12}} \frac{kT}{P} \left[\frac{2kT(m_1 + m_2)}{\pi m_1 m_2} \right]^{\frac{1}{2}},$$

where m_1 and m_2 are the respective masses of 'air molecules' and hydrocarbon, P is the pressure, D_{12} the corresponding diffusion coefficient, and α_{12} a factor, the inclusion of which is not wholly satisfactory (cf. Furry 1948), to allow for the persistence of velocities. From Jeans's data α_{12} is 0.304, 0.306 and 0.308 for the hydrocarbons *n*-C₁₆H₃₄, *n*-C₁₇H₃₆ and *n*-C₁₈H₃₈ respectively. By subtracting from S_{12} the radius of an 'air molecule', 1.87 Å (Jeans), we obtain the values of the radii given in table 6. This assumes that the 'air molecules' have a constant radius over the small temperature range adopted, which is very nearly true. Allowance for the temperature coefficient of the radius of 'air molecules', as deduced from the Sutherland constant for air, would affect the figures for the higher temperature by only 0.5 %, and we have not included this correction because the Sutherland constant is deduced from experiments in which collisions between like molecules occur, whereas in our experiments 'air molecules' collide with hydrocarbon molecules. The last column in table 6 refers to the discussion below.

TABLE 6. RADII OF HYDROCARBON MOLECULES DEDUCED FROM COLLISION FORMULAE ($\frac{1}{2}S_{11}$)

temp. (°C)	<i>n</i> -C ₁₆ H ₃₄	<i>n</i> -C ₁₇ H ₃₆	<i>n</i> -C ₁₈ H ₃₈	$\left(\frac{1+\alpha'}{1-\alpha'}\right)^{\frac{1}{2}}$
14.97	4.53	4.40	4.56	1.705
16.97	4.52	—	—	1.697
20.01	4.51	4.37	4.55	1.686
24.98	4.49	4.34	4.53	1.668
29.98	4.46	4.31	4.50	1.652
34.88	4.22	4.29	4.47	1.636
39.94	—	4.27	4.44	1.620

The variation with temperature of the collision radius is another aspect of the temperature dependence of D_{12} , already discussed empirically. According to Sutherland this variation may be represented by

$$S_{12}^2 = S_{\infty}^2(1 + C/T),$$

where C is the Sutherland constant and S_{∞} the value of S_{12} at very high temperatures (the formula clearly does not apply to very low temperatures). Sutherland's formula would correspond to a variation in D_{12} of the type $D \propto T^{\frac{1}{2}}/(T + C)$. The Sutherland constants for the hydrocarbons have been calculated by plotting $T^{\frac{1}{2}}/D_{12}$ against T , and are 560, 590 and 548 for $n\text{-C}_{16}\text{H}_{34}$, $n\text{-C}_{17}\text{H}_{36}$ and $n\text{-C}_{18}\text{H}_{38}$ respectively. These values are roughly the same, and may be compared with those of Melaven & Mack (1932), who found from viscosity measurements on $n\text{-C}_7\text{H}_{16}$, C_8H_{18} and C_9H_{20} the respective values 445, 337 and 276 for the Sutherland constants. All that can be said from this point of view is that our values of C are of the right order, since Sutherland's constants were developed to cover collisions of like molecules, the case studied by Melaven & Mack. It should also be noticed that an analysis in terms of Sutherland's formulae was not possible for di- n -butyl phthalate (cf. part II), since the temperature variation of D_{12} was too great.

The temperature coefficient of D_{12} and S_{12} may also be treated in terms of repulsive and attractive forces (cf. part II), since the variation is within the maximum allowed for simple molecules, and is also in accord with the observations of von Obermayer (1880), who found for experiments involving collisions between molecules of the more condensable gases and vapours, e.g. water and carbon dioxide, $D_{12} \propto T^2$. It is, however, somewhat unreal to treat a flexible molecule such as a hydrocarbon by the same methods as for simple molecules, and a more realistic treatment is given below.

The values of the radii given above are in reasonable accord with those deduced from the density of the liquid, assuming close packing of spherical molecules as a very rough approximation to the true packing; thus for $n\text{-C}_{16}\text{H}_{34}$ this assumption gives a value 4.41 Å for the radius at 20° C, in comparison with 4.51 Å deduced from the diffusion coefficient. Our results are not in agreement, however, with the empirical formula of Burk, Laskowski & Lankelma (1941), based on the data of Titani (1930), and of Melaven & Mack (1932) on the lower hydrocarbons up to nonane. The equation of Burk *et al.* for the radii of hydrocarbon molecules, viz.

$$\frac{1}{2}S_{11} = 0.23(n - 2) + 1.84 \quad (\text{in Å}),$$

where n is the number of carbon atoms in the molecule, gives for

$$n\text{-C}_{16}\text{H}_{34} \quad \frac{1}{2}S_{11} = 5.06 \text{ Å}, \quad \text{and for} \quad n\text{-C}_{18}\text{H}_{38} \quad \frac{1}{2}S_{11} = 5.52 \text{ Å},$$

considerably greater than the experimental values. Moreover, the equation indicates that the radius increases by 0.23 Å per CH_2 group, which is not in accordance with our results and is unlikely on general grounds for the higher hydrocarbons (*vide infra*).

For the solid hydrocarbons the work of Müller (1925-41) shows that the hydrocarbon molecules are arranged as parallel zigzag chains, thermal oscillation being insufficient to overcome the intermolecular forces. The main influence of increasing

temperature is to separate the 'sideways' distance between the chains. For the liquid state the evidence is less conclusive, although the X-ray studies of Stewart (1928) and of Warren (1933) suggest that there is considerable adlineation of the molecules. It is probable, however, that the increased distance between the chains, and the consequent weaker intermolecular forces, allows for some degree of statistical coiling to occur; *intramolecular* forces will, however, be small.

In the vapour phase, however, this statistical coiling will be the predominant factor affecting the configuration. As is well known very long flexible hydrocarbon chains with free rotation about the links will assume the configuration of greatest entropy, given by $\overline{L^2} = l^2 \left(\frac{1 + \cos \theta}{1 - \cos \theta} \right) z$, where $\overline{L^2}$ is the mean square of the distance between the centres of the end-carbon atoms, z is the number of links (i.e. $z = n - 1$), l is the length of a link, and θ is the supplement of the valency angle. For shorter chains Eyring (1932) developed an exact formula,

$$\overline{L^2} = l^2 [z + 2(z-1) \cos \theta + 2(z-2) \cos^2 \theta + \dots + 2 \cos^{z-1} \theta].$$

A direct summation for *n*-octadecane gives $\overline{L^2} = 1.54^2 \times 33.09 \text{ \AA}^2$, i.e. $(\overline{L^2})^{\frac{1}{2}} = 8.86 \text{ \AA}$, instead of $(\overline{L^2})^{\frac{1}{2}} = 8.98 \text{ \AA}$, the value for octadecane if the long chain formula were applied. The difference in $(\overline{L^2})^{\frac{1}{2}}$, amounting to 1 %, is very small.

With any given value of L a large number of configurations are possible, and the value of L gives only a rough guide to the mean molecular size, although it is clear from the above result for octadecane, when compared with the length of the molecule, that considerable coiling must occur in the vapour purely for statistical reasons. A more useful parameter for comparison with our results is the mean square of the distance between carbon atom centres and the mass centre ($\overline{R^2}$). Taylor (1948) writes for long chains and for a summation over all molecular configurations,

$$\overline{R^2} = \frac{1}{8}(\overline{L^2}), \quad \text{or} \quad (\overline{R^2})^{\frac{1}{2}} = 0.4083(\overline{L^2})^{\frac{1}{2}}.$$

For octadecane this gives $(\overline{R^2})^{\frac{1}{2}} = 3.67 \text{ \AA}$, if we assume that the formula for long chains applies. It should be noted that this result refers to distances from the mass centre to the *centres* of carbon atoms.

This value of $(\overline{R^2})^{\frac{1}{2}}$ is considerably less than the experimental collision radius at 20° C, viz. 4.55 Å; no allowance has been made, however, for the van der Waals radius of the CH₂ group in the value 3.67 Å for $(\overline{R^2})^{\frac{1}{2}}$. Moreover, the temperature dependence of $(\overline{R^2})^{\frac{1}{2}}$ is small, being confined to the variation in l . It is possible that no direct comparison can be made between $(\overline{R^2})^{\frac{1}{2}}$ and the radius found from collision formulae. There is a second possibility, however, namely, that the rotation about a bond is not free. Thermodynamic studies on the lower hydrocarbons have proved that energy barriers to free rotation exist, and this work has been extended to longer chains by Taylor (1948). Taylor finds that for long chains

$$\overline{L^2} = l^2 \left(\frac{1 + \cos \theta}{1 - \cos \theta} \right) \left(\frac{1 + \alpha'}{1 - \alpha'} \right) z, \quad \overline{R^2} = \frac{1}{8} \overline{L^2},$$

where

$$\alpha' = \int_0^\pi e^{-V(\phi)/(RT)} \cos \phi d\phi / \int_0^\pi e^{-V(\phi)/(RT)} d\phi,$$

$V(\phi)$ being the potential energy for rotation about the $z-2$ interior bonds, the zero of the angle of rotation of ϕ being in the *trans* position. Taylor writes

$$V(\phi) = \frac{1}{2}V_m[x(1 - \cos\phi) + (1-x)(1 - \cos 3\phi)],$$

where V_m is the maximum value of V and the factors x and $(1-x)$ allow for the contributions of one- and threefold components respectively. From the thermodynamic data for the lower hydrocarbons $V_m = 4100$ cal./mole, and $x = 0.26$; the factor $\left(\frac{1+a'}{1-a'}\right)^{\frac{1}{2}}$ has, according to Taylor, the following values: -50°C 2.02, 0°C 1.76, 50°C 1.59, 100°C 1.48, 150°C 1.41, 200°C 1.36, 300°C 1.29, 400°C 1.23. There is thus a considerable temperature dependence of \bar{R}^2 , due to the change in statistical configuration, quite apart from any consideration of decreasing radius due to increasing velocity of impact on increasing the temperature. We have interpolated graphically the values of $[(1+a')/(1-a')]^{\frac{1}{2}}$ given in table 6, and find that our results on the radius deduced from collision formulae are considerably less than

$$l \left[\frac{1}{6} \left(\frac{1 + \cos\theta}{1 - \cos\theta} \right) \left(\frac{1 + a'}{1 - a'} \right) z \right]^{\frac{1}{2}}.$$

Thus, for octadecane at 20°C Taylor's formula gives for $(\bar{R}^2)^{\frac{1}{2}}$ the value 6.19 \AA , and at 35°C the value 6.00 \AA . If allowance is made for the van der Waals radius of CH_2 the discrepancy will be greater.

On the assumption of a Gaussian error distribution for a long chain with free rotation, the most probable and the mean values of L may be deduced from the mean square values. Thus the mean value of $R = 3.36\text{ \AA}$ for octadecane. The various results we have obtained may be summarized (making a rough allowance for the van der Waals radius of CH_2):

$(\bar{R}^2)^{\frac{1}{2}}$ + van der Waals's radius of $\text{CH}_2 = 5.67\text{ \AA}$; free rotation.

\bar{R} + van der Waals's radius of $\text{CH}_2 = 5.36\text{ \AA}$; free rotation.

$(\bar{R}^2)^{\frac{1}{2}}$ + van der Waals's radius of $\text{CH}_2 = 8.19\text{ \AA}$; restricted rotation.

Our results would not be greatly changed if we had used the first approximation for rigid elastic spheres of Chapman & Cowling (1939) for the calculation of S_{12} ; the use of this approximation would involve increasing our values of S_{12} by 7 %. Higher approximations would involve changes of only a few per cent. Although no identification of the experimental radius with the above radii is possible on theoretical grounds, it would have been expected that the theoretical result would be closer to the experimental value than is observed to be the case for restricted rotation. In this connexion it should be remembered that *intramolecular* forces, which will favour coiling, are neglected in Taylor's theory.

Our results seem to be best interpreted on the hypothesis of a helical molecule, as a representation of the mean position, with the ends at the statistical distance apart, for free rotation. We have constructed such a model from units kindly lent by Professor M. G. Evans, the units being moulded to give the correct relative van der Waals radii in the crystal. It will be seen from figure 1, that a coiled helix is possible for *n*-octadecane, the molecule assuming an approximately spherical form.

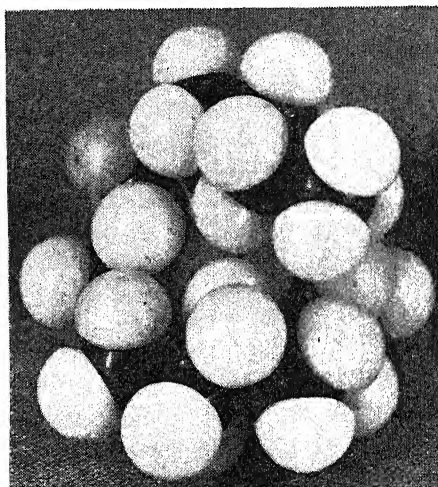


FIGURE 1. Coiled form of $C_{18}H_{33}$

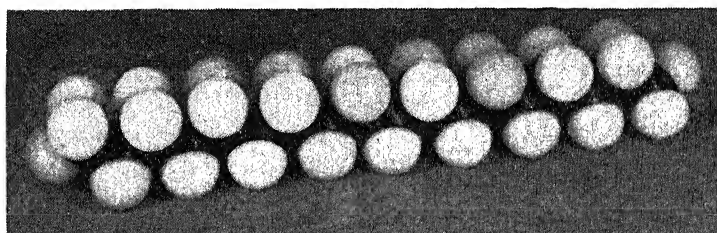


FIGURE 2. Straight chain form of $C_{18}H_{38}$

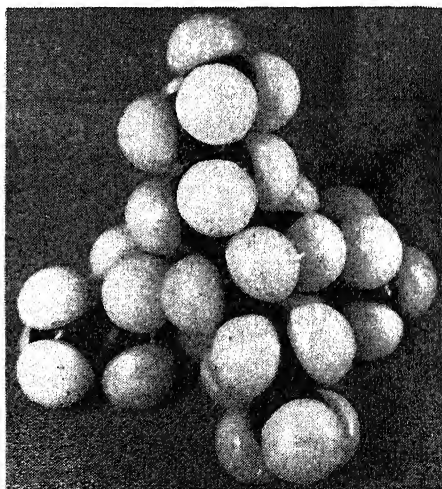


FIGURE 3. Tetra-*n*-butyl methane

This is in agreement with the close values of the radii of $C_{16}H_{34}$, $C_{17}H_{36}$ and $C_{18}H_{38}$, since a similar representation is possible for the two lower hydrocarbons. For comparison figure 2 and 3 give the straight-chain form of octadecane, and the branched-chain molecule $C_{17}H_{36}$, tetra-*n*-butyl methane. The hypothesis of coiling had previously been advanced by Mack and his co-workers to explain the comparatively low values of the collision radii of the lower hydrocarbons, although with hydrocarbons up to C_9H_{20} the 'helix', on our model, would occupy scarcely one turn. When the theory of Taylor is applied to these lower hydrocarbons discrepancies appear similar to those discussed above.

Our thanks are due to the Shell Refining and Marketing Company, Ltd., for a grant and for generously allowing one of us (A.D.S.) to take up the work at Leeds, in their employ, and to submit the work in partial fulfilment of the requirements for a Ph.D. degree at the University of Leeds. We are also grateful to Imperial Chemical Industries, who gave a grant for chemicals, and to Professor M. G. Evans, F.R.S., for his constant interest and encouragement.

REFERENCES

- Birks, J. & Bradley, R. S. 1949 *Proc. Roy. Soc. A*, **198**, 226.
Bradley, R. S., Evans, M. G. & Whytlaw-Gray, R. W. 1946 *Proc. Roy. Soc. A*, **186**, 368.
Burk, R., Laskowski, L. & Lankelma, H. P. 1941 *J. Amer. Chem. Soc.* **63**, 3248.
Carey, P. C. & Smith, J. C. 1933 *J. Chem. Soc.* pp. 346, 1348.
Chapman, S. & Cowling, T. G. 1939 *Mathematical theory of non-uniform gases*. Cambridge University Press.
Eyring, H. 1932 *Phys. Rev.* **39**, 746.
Fuchs, N. 1934 *Phys. Z. Sowjet.* **6**, 225.
Furry, W. H. 1948 *Amer. J. Phys.* **16**, 163.
Huggins, M. 1939 *J. Phys. Chem.* **43**, 1083.
Jeans, J. 1940 *An introduction to the kinetic theory of gases*. Cambridge University Press.
Knudsen, M. 1909 *Ann. Phys., Lpz.*, **29**, 179.
Melaven, R. M. & Mack, E. 1932 *J. Amer. Chem. Soc.* **54**, 888.
Müller, A. & Saville, W. B. 1925 *J. Chem. Soc.* **127**, 599.
Müller, A. 1927 *Proc. Roy. Soc. A*, **124**, 317.
Müller, A. 1930 *Proc. Roy. Soc. A*, **127**, 417.
Müller, A. 1932 *Proc. Roy. Soc. A*, **138**, 514.
Müller, A. 1936 *Proc. Roy. Soc. A*, **154**, 624.
Müller, A. 1941 *Proc. Roy. Soc. A*, **178**, 227.
von Obermayer 1880 *S.B. Akad. Wiss. Wien.* **81**, 1102.
Rossini, F. D. 1945 *Bur. Stand. J. Res., Wash.*, **34**, 263.
Stewart, J. 1928 *Phys. Rev.* **31**, 174.
Taylor, W. J. 1948 *J. Chem. Phys.* **16**, 257.
Titani, T. 1929 *Bull. Chem. Soc. Japan*, **4**, 277.
Titani, T. 1930 *Bull. Chem. Soc. Japan*, **5**, 98.
Ubbelohde, A. R. 1938 *Trans. Faraday Soc.* **34**, 282.
Warren, B. W. 1933 *Phys. Rev.* **44**, 969.
Zilberman-Granovskaja, A. A. & Schugan, E. A. 1940 *J. Phys. Chem. U.S.S.R.* **14**, 759.

The autoxidation of tetralin

BY C. H. BAMFORD AND M. J. S. DEWAR

Courtaulds Ltd., Maidenhead, Berks

(Communicated by A. H. Wilson, F.R.S.—Received 28 December 1948)

The autoxidation of tetralin can be photosensitized by suitable vat dyes. This reaction can therefore be investigated by standard photochemical techniques. The kinetics have in this way been shown to conform to the scheme suggested by Bolland & Gee (1946) for autoxidations. The absolute velocity constants for the elementary processes have been evaluated, and the activation energies and frequency factors estimated (table 4).

The frequency factors of free radical reactions in solution have been discussed. It is suggested that the 'normal' frequency factor for such a reaction is about 10^7 .

It is pointed out that the techniques described in this paper should be generally applicable to the study of free radical reactions in solution.

INTRODUCTION

The autoxidation of tetralin has previously been studied by George, Rideal & Robertson (1946), George & Robertson (1946*a, b*), George (1946*a, b*) and Robertson & Waters (1946). These authors now seem to agree that the oxidation is a radical chain process, following the general scheme of Bolland & Gee (1946), although the evidence that the reaction does follow these kinetics is somewhat scanty.

We have recently found that many vat dyes are able to photosensitize the autoxidation of tetralin (Bamford & Dewar 1948*a*). By initiating the reaction in this way it is possible to make use of standard photochemical techniques, and we have thus been able to confirm the general mechanism, and also to determine the absolute values of the velocity constants of the individual radical reactions.

EXPERIMENTAL METHOD

The reactions were carried out in a Warburg type apparatus, and pressure changes were followed by means of a dibutyl phthalate manometer read to 0.01 mm. with a cathetometer. The consumption of oxygen could be calculated from the pressure change and the dimensions of the apparatus. The reaction vessel was a spherical bulb of approximately 20 ml. capacity, attached to the manometer by a glass spiral, and shaken mechanically at constant speed. The rate of shaking was normally about 7 times per sec., and was measured stroboscopically. Five ml. of tetralin were used in each run. The whole apparatus was immersed in a transparent thermostat ($\pm 0.01^\circ \text{C}$). The system could be evacuated to 10^{-5} mm. Hg by a three-stage mercury-diffusion pump backed by an oil pump and could be filled with oxygen (previously purified by fractional distillation) at pressures measured on mercury or Apiezon oil manometers.

In the photochemical experiments a 1 kW projection lamp was used as light source. A 2 l. flask filled with 10 % copper sulphate solution was used as a condensing lens, and served also to remove infra-red radiation. The solution was

circulated by a pump to avoid overheating. The rest of the system consisted of an iris diaphragm, an electromagnetically operated Compur shutter, and two collimating lenses. Filters of Crookes A glass and saturated sodium nitrite solution could be inserted. Intensities were measured by a calibrated thermopile potentiometer system. Normally both filters were used, cutting off light with $\lambda < 4200 \text{ \AA}$.

Purification of materials

Tetralin was shaken with concentrated sulphuric acid, and after washing and drying it was boiled under reflux for several hours with metallic sodium, in a nitrogen atmosphere. It was then distilled under nitrogen through a 15-plate column, and preserved in evacuated ampoules.

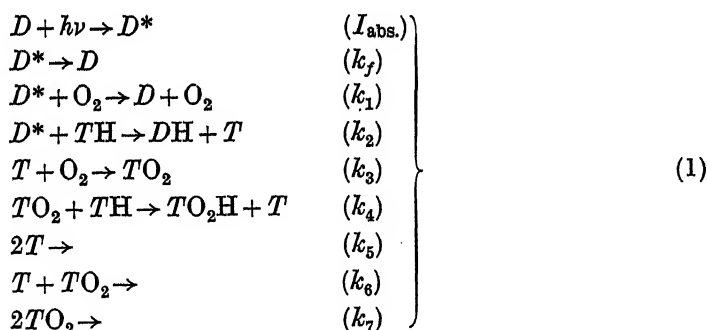
The dyes, specially pure specimens of which were kindly supplied by I.C.I., were further purified by sublimation *in vacuo*. Solutions of the dyes in tetralin (0.05 mg./ml.) were made up in evacuated tubes.

Benzoyl peroxide was crystallized from chloroform and light petroleum, and dried *in vacuo*.

Hydroquinone, and α - and β -naphthols were pure commercial specimens.

KINETIC SCHEME

It is convenient to give the complete kinetic scheme at this stage. This differs from the scheme of Bolland & Gee only in the mode of initiation of the chains:



D and TH represent dye and tetralin respectively; T is a radical formed from tetralin by loss of a hydrogen atom, and TO_2 is a peroxide radical derived from T . For long chains, with the approximation $k_6 = \sqrt{(k_5 k_7)}$ (Bolland & Gee 1946), this scheme leads to

$$-\frac{d[O_2]}{dt} = \left\{ \frac{k_2 I_{\text{abs.}} [TH]}{k_f + k_1 [O_2]_l + k_2 [TH]} \right\}^{\frac{1}{2}} \frac{k_3 k_4 [O_2]_l [TH]}{k_3 \sqrt{k_7} [O_2]_l + k_4 \sqrt{k_5} [TH]}, \quad (2)$$

where $[O_2]_l$ is the concentration of oxygen in the liquid. The term in the first bracket on the right of (2) gives the rate of chain starting I . These equations assume that the light absorption is weak. It was established that this was so in the present experiments.

RESULTS AT HIGH OXYGEN PRESSURES

At sufficiently high oxygen pressures the second term on the right of (2) becomes effectively constant. If the rate of initiation is constant the rate of reaction should then be independent of the oxygen pressure. Initiation by benzoyl peroxide should fulfil this condition (cf. Bolland 1948). Table 1 shows that in fact the rate of reaction remained sensibly constant down to pressures of about 50 mm. At low pressures the rate fell as equation (2) requires; the dependence of the rate on the pressure will be discussed later.

TABLE 1. AUTOXIDATION OF TETRALIN INITIATED BY BENZOYL PEROXIDE

$T = 25^\circ \text{C}$, benzoyl peroxide, 1%		$T = 45^\circ \text{C}$, benzoyl peroxide, 0.1%	
oxygen pressure (mm. Hg)	$-10^7 d[\text{O}_2]/dt$ (mol.l. ⁻¹ sec. ⁻¹)	oxygen pressure (mm. Hg)	$-10^7 d[\text{O}_2]/dt$ (mol.l. ⁻¹ sec. ⁻¹)
600	3.09	600	5.98
400	3.04	400	5.97
300	3.04	200	5.95
200	3.01	100	5.98
50	2.98	5.42	5.64
10	2.91	2.89	5.30
1.93	2.72	1.45	4.57
1.55	2.64	1.30	4.31
1.16	2.53	0.99	4.02
0.772	2.30	0.685	3.37
0.618	2.07	0.640	3.03
0.463	1.85	0.535	2.77
0.386	1.68	0.383	2.14
0.309	1.48	0.283	1.73
0.231	1.21		
0.154	0.86		
0.116	0.66		
0.077	0.47		
0.039	0.25		

When the reaction is photosensitized by dyes, the rate for a given dye and light intensity is not constant at high oxygen pressures, but falls off as the pressure increases. This indicates that the rate of initiation must depend on the pressure. It is known that excited dyes are deactivated by oxygen. The second and third equations in scheme (1) allow for spontaneous (or solvent) deactivation and deactivation by oxygen respectively. From equation (2) it follows that at high oxygen pressures

$$\left(\frac{d[\text{O}_2]}{dt}\right)^{-2} = \frac{k_7}{k_4^2 k_2 I_{\text{abs.}} [\text{TH}]^3} \{k_f + k_1 [\text{O}_2] + k_2 [\text{TH}]\}. \quad (3)$$

Assuming that solutions of oxygen in tetralin obey Henry's law a plot of $(d[\text{O}_2]/dt)^{-2}$ against pressure should be linear, and the ratio of slope to intercept gives the 'relative quenching coefficient' q for oxygen, defined by

$$q = \frac{k_1}{k_f + k_2 [\text{TH}]}. \quad (4)$$

Figure 1 shows that (3) holds for Caledon Red BN and Caledon Golden Yellow GK, the values of q at 25° C being 2.89×10^2 and 1.27×10^3 respectively with the oxygen concentration in moles/litre. Throughout this paper the absorption coefficient of oxygen in tetralin is taken to be 0.10 (Fischer & Pfeiderer 1922), and the heat of solution has been neglected. The value of q for Red BN changes little with temperature, the value at 0° C estimated from figure 1 being 3.64×10^2 .

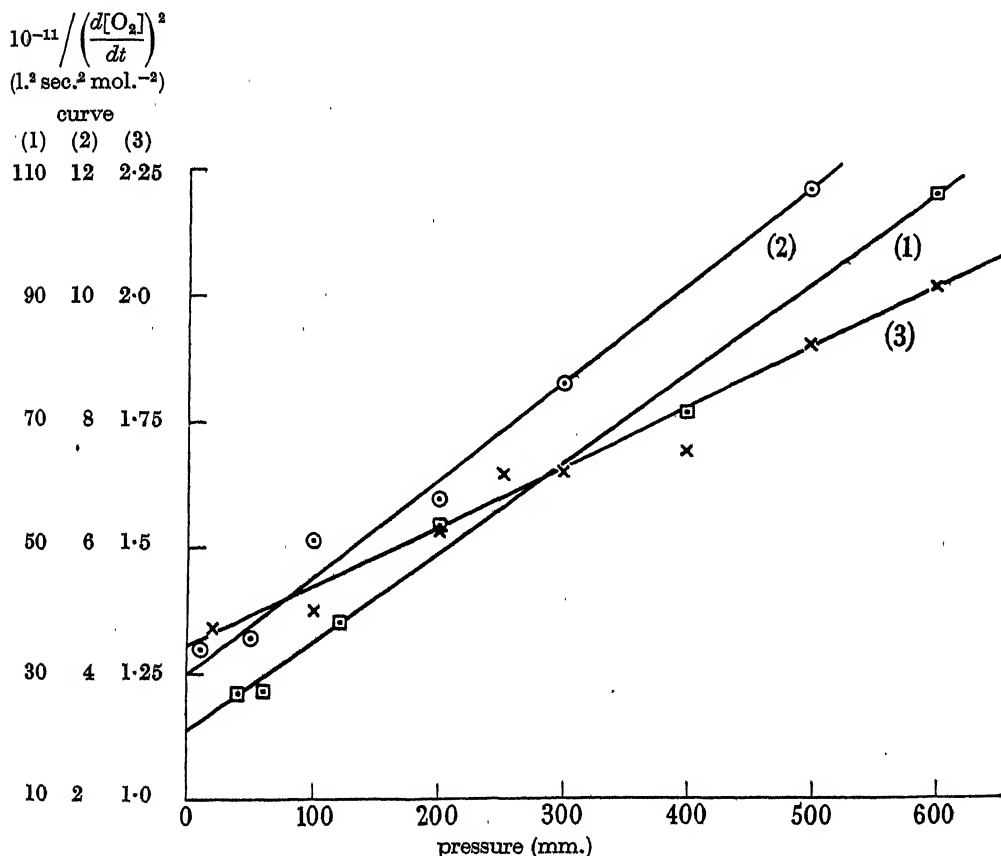


FIGURE 1. Determination of quenching. Curve (1), \square , Caledon Golden Yellow GK, 25° C. Curve (2), \circ , Caledon Red BN, 25° C. Curve (3), \times , Caledon Red BN, 0° C.

Effect of intensity

According to equation (2), the rate of reaction should be proportional to the square root of the intensity at constant oxygen pressure. A plot of $\log_{10} I$ against $\log_{10} (\text{rate})$ is given in figure 2. The intensity exponent is 0.55; the small discrepancy is probably due to the non-uniformity of the light beam across which the reaction vessel is shaken, and which would vary with the setting of the iris diaphragm.

Measurement of the rate of chain starting

In order to determine the absolute velocity constants it is necessary to know the rate of chain starting. This can conveniently be measured by an inhibitor method. Bolland & ten Have (1947 *a, b*) have shown that in the autoxidation of ethyl linoleate

inhibited by phenols, only the peroxide radicals are removed by the inhibitor. Bolland & ten Have could not obtain a reliable value for the rate of chain starting, since in the peroxide-initiated reaction the nature of the initial radicals is uncertain. In our case, however, with photochemical initiation, the initial radicals are tetralyl, and each radical must react with oxygen before it can be removed by the inhibitor. The rate of oxygen uptake extrapolated to infinite inhibitor concentration should

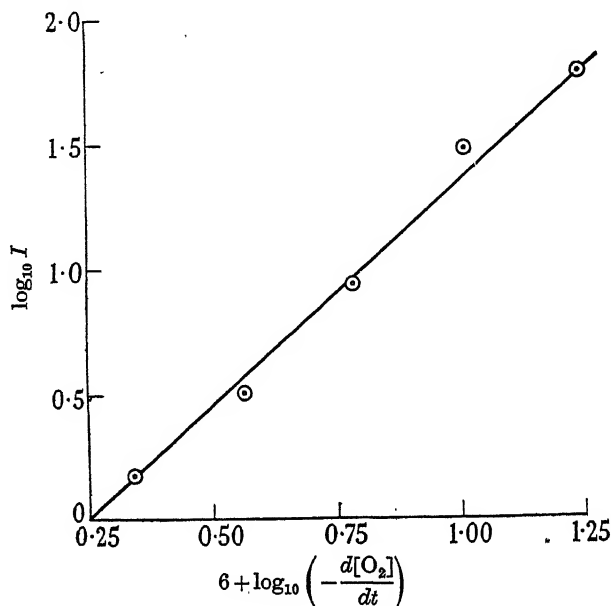


FIGURE 2. Dependence of rate on light intensity.

then equal the rate of chain starting. The reaction below (velocity constant k_8) must be included in the kinetic scheme



For high inhibitor and oxygen concentrations this leads to

$$-\frac{d[O_2]}{dt} = I \left(1 + \frac{k_4[TH]}{k_8[P]} \right). \quad (6)$$

Under these conditions a plot of rate against $1/[P]$ should be a straight line, and the intercept corresponding to $1/[P] = 0$ should be independent of the nature of the inhibitor and equal to the rate of chain starting. Figure 3 shows that equation (6) holds for reactions initiated by Cibanone Yellow 2GR and inhibited by α - and β -naphthols and hydroquinone. The fact that the intercept ($2.50 \times 10^{-7} \text{ mol.l.}^{-1}$) is not zero provides confirmation that the inhibitors do not react appreciably with the tetralyl radicals. The rate of reaction in the absence of inhibitors, under the same conditions was $1.055 \times 10^{-5} \text{ mol.l.}^{-1}\text{sec.}^{-1}$. Hence at 25°C

$$I = 2.25 \times 10^3 (d[O_2]/dt)^2. \quad (7)$$

Similar measurements at 45°C lead to the relation

$$I = 0.915 \times 10^3 (d[O_2]/dt)^2. \quad (8)$$

Quenching of the excited dye by oxygen is irrelevant in these measurements, since all the experiments were carried out at the same oxygen pressure (600 mm.). It is, however, assumed that the phenol does not deactivate the excited dye significantly. The results of Weber (1948) and the results with the different inhibitors shown in figure 3, suggest that this is so under our conditions.

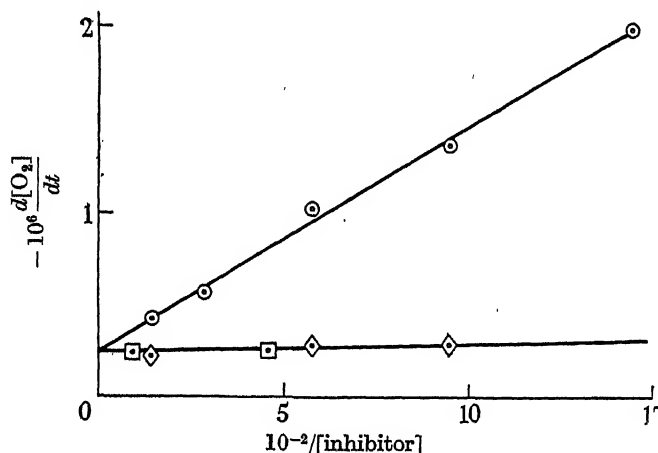


FIGURE 3. Determination of rate of inhibition.
 \circ β -naphthol, \square hydroquinone, \diamond α -naphthol

At high oxygen pressures equation (2) reduces to

$$-\frac{1}{\sqrt{I}} \frac{d[O_2]}{dt} = \frac{k_4[TH]}{\sqrt{k_7}} \quad (9)$$

From (8) and (9) and with $[TH] = 7.36 \text{ mol.l.}^{-1}$,

$$\left. \begin{aligned} k_4/\sqrt{k_7} &= 2.87 \times 10^{-3} \text{ l.}^{1/2} \text{ mol.}^{-1/2} \text{ sec.}^{-1} \text{ at } 25^\circ \text{ C.} \\ k_4/\sqrt{k_7} &= 4.50 \times 10^{-3} \text{ l.}^{1/2} \text{ mol.}^{-1/2} \text{ sec.}^{-1} \text{ at } 45^\circ \text{ C.} \end{aligned} \right\} \quad (10)$$

Thus $E_4 - \frac{1}{2}E_7 = 4.3 \text{ kcal.}$

Incidentally, this technique cannot be used for reactions initiated by benzoyl peroxide. First, the nature of the initial radicals is uncertain; benzoate radicals would be removed by the inhibitor. Secondly, if phenyl radicals were formed, the concomitant evolution of carbon dioxide would interfere, and this effect would be augmented by any primary recombination of phenyl and benzoate radicals. In attempting to measure rates of chain starting by this method, we have in fact observed pressure increases, which imply that the rate of evolution of carbon dioxide is faster than the absorption of oxygen.

Absolute determination of k_4 and k_7 by the rotating sector method

From (1) it follows that the mean lifetime τ of the kinetic chain is given by

$$\tau = \frac{1}{\sqrt{I} \frac{k_3[O_2]_l + k_4[TH]}{k_3\sqrt{k_7}[O_2]_l + k_4\sqrt{k_8}[TH]}} \quad (11)$$

At high oxygen pressures this reduces to

$$\tau = 1/\sqrt{(k_7 I)}. \quad (12)$$

Thus the determination of τ at high oxygen pressures enables k_4 and k_7 to be calculated absolutely from equations (10) and (12).

Measurement of the lifetime was carried out by the rotating sector method of Briers, Chapman & Walters (1926), using a 1:1 ratio. Since periods of interruption of the order of seconds were required, a shutter was used in place of a rotating sector. The shutter was operated by a commutator driven by a synchronous motor. The results

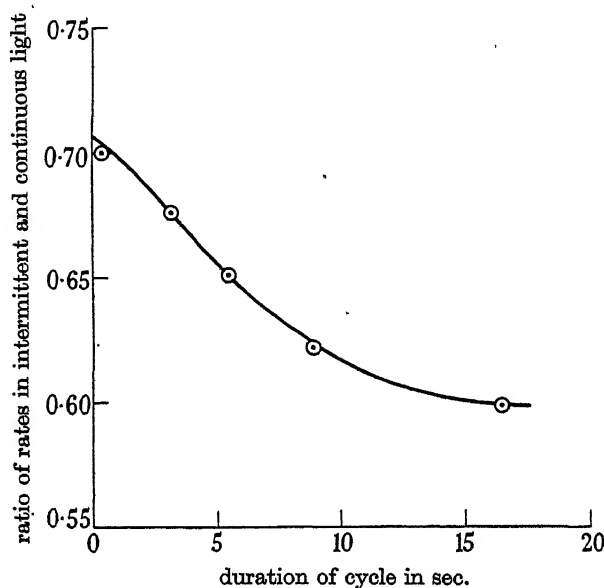


FIGURE 4. Determination of lifetime by the sector method.

were analyzed in the manner recommended by Dickinson (see Noyes & Leighton 1939) and followed the theoretical curve within the limits of experimental error (see figure 4). Further, the lifetimes were proportional to $I^{-\frac{1}{2}}$ (table 1), showing that the life of the excited dye was negligible compared to that of the chain (otherwise the measured lifetimes would be given by $\tau + \text{constant}$).

TABLE 2. DEPENDENCE OF LIFETIMES ON INTENSITY

$-d[\text{O}_2]/dt \times 10^6$ mol. l. ⁻¹ sec. ⁻¹	$I^{\frac{1}{2}} \times 10^4$ (mol. l. ⁻¹ sec. ⁻¹) ^{$\frac{1}{2}$}	τ (sec.)	$\tau I^{\frac{1}{2}} \times 10^4$
12.95	6.14	0.24	1.47
4.03	1.96	0.86	1.69
1.61	0.76	2.03	1.54

From table 2 and equations (10) and (12) we have

$$k_4 = 18.3 \text{ l. mol.}^{-1} \text{ sec.}^{-1}, \quad k_7 = 4.1 \times 10^7 \text{ l. mol.}^{-1} \text{ sec.}^{-1}. \quad (13)$$

RESULTS AT LOW OXYGEN PRESSURES

According to equation (2), if $[O_2]_l$ is always proportional to the concentration of oxygen in the gas phase, $(O_2)_g$,

$$-1 \left/ \frac{d[O_2]_l}{dt} \right. = a + \frac{b}{(O_2)_g}, \quad (14)$$

where a, b are constants for a given rate of chain starting. The experimental results, however, do not fit equation (14), as will be seen from figure 5. The rate falls off too rapidly at low oxygen pressures. Thus either the kinetic scheme (1) must be wrong or the solution of oxygen must be rate determining at low oxygen pressures. The latter explanation was shown to be correct by experiments with interrupted light.

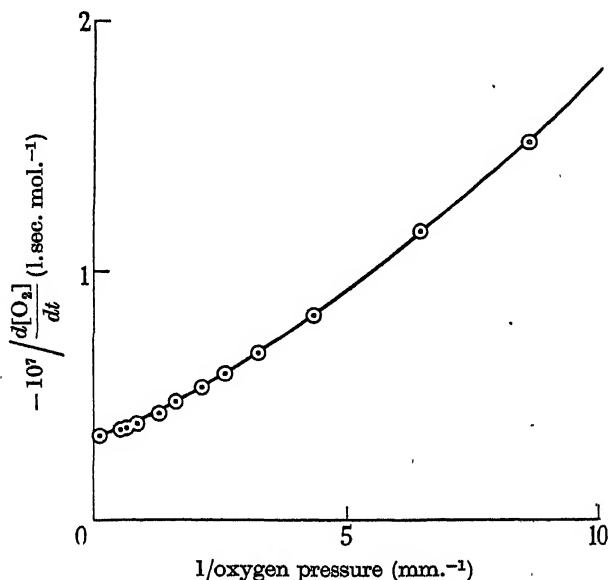


FIGURE 5. Benzoyl peroxide 1%, 25° C.

At low oxygen pressures the rate was much the same in continuous and intermittent light, a fact which can be understood only if solution of oxygen is rate-determining. It may be assumed that the rate of solution of oxygen is proportional to the difference between the saturation concentration and the actual concentration in the liquid, i.e. that

$$-d\{O_2\}/dt = k\{s(O_2)_g - [O_2]_l\}, \quad (15)$$

where k is constant at constant temperature and constant rate of shaking, s is the absorption coefficient, and $d\{O_2\}/dt$ represents the rate of disappearance of oxygen from the gas phase, in moles per litre of solution per second. From (14) and (15) it is easy to show that

$$-(O_2)_g s \left\{ a + 1 \left/ \frac{d\{O_2\}}{dt} \right. \right\} = b + \frac{1}{k} + \frac{a}{k} \frac{d\{O_2\}}{dt}. \quad (16)$$

It can be seen from (14) that a is the limiting value of $-1 \left/ \frac{d\{O_2\}}{dt} \right.$ at high oxygen pressures. Thus the left-hand side of equation (16) can be calculated in any experi-

ment, and, by plotting it against $-d\{O_2\}/dt$, a straight line should be obtained from which b and k can be found. Figures 6 and 7 show this plot for thermal initiation by benzoyl peroxide and for photo-initiation by Caledon Red BN at 25°. The experimental points lie satisfactorily on straight lines. From (2),

$$b = \sqrt{k_5/(k_3 \sqrt{I})}, \quad (17)$$

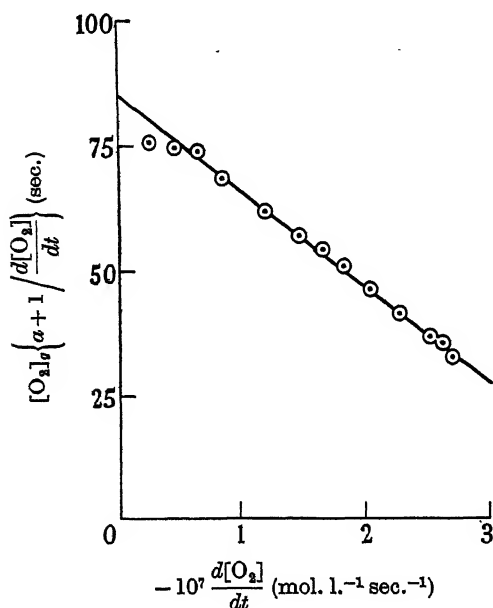


FIGURE 6. Benzoyl peroxide, 1%, 25° C
($I = 1.98 \times 10^{-10}$).

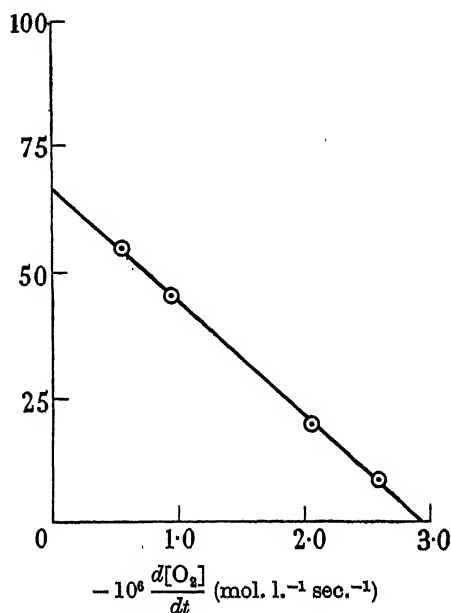


FIGURE 7. Caledon Red BN, 25° C
($I = 2.25 \times 10^{-8}$).

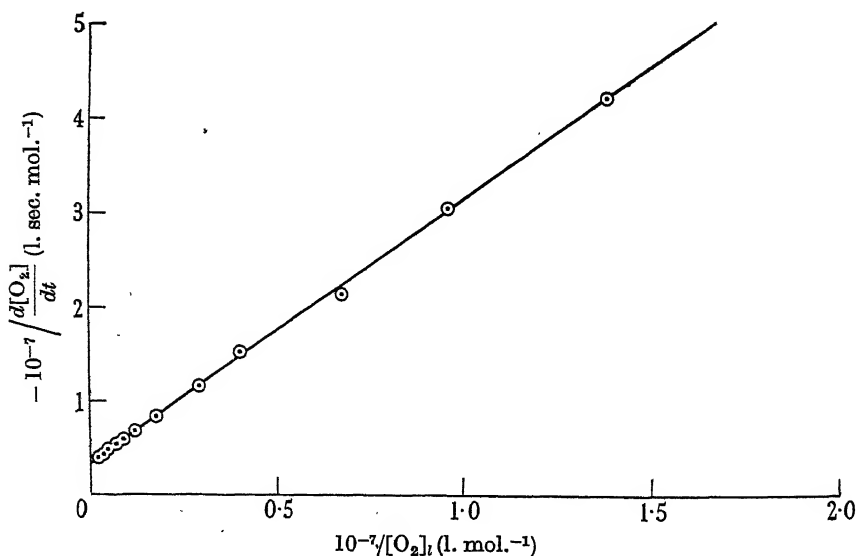


FIGURE 8. Benzoyl peroxide 1%, 25° C.

and $b\sqrt{I}$ should be constant. The values obtained for this quantity are 3.94×10^{-5} (figure 6) and 3.90×10^{-5} (figure 7). We may take the value of $\sqrt{k_5/k_3}$ to be 3.94×10^{-5} at 25°C , since figure 6 represents the more accurate series of experiments. Using the value of k from figure 6, values of $[\text{O}_2]_t$ may be calculated from (15). According to equation (14) a plot of $-1 \left/ \frac{d[\text{O}_2]}{dt} \right.$ against $1/[\text{O}_2]_t$ should be a straight line. Figure 8 shows that this is so. Although the straight lines in figures 6 and 8 are consequences of the same equations, we consider that b can be found more conveniently from figure 8. In this case the value agrees almost exactly with that obtained from figure 6. It will be noticed that the magnitudes of k found from figures 6 and 7 are not quite the same. The value of k should be sensitive to the experimental conditions, e.g. the exact position of the reaction bulb, and much greater variations were obtained by altering the rate of shaking.

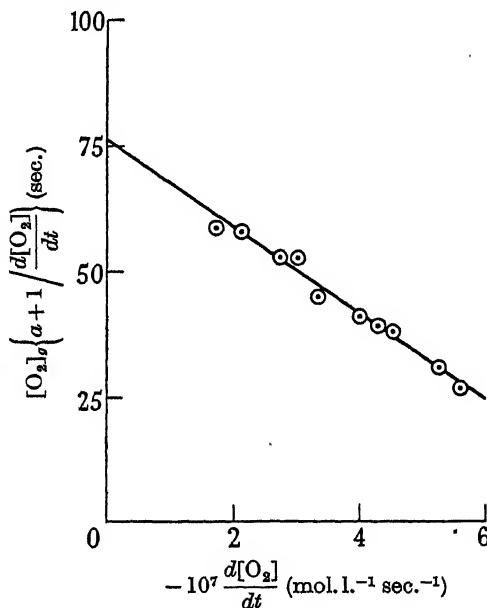


FIGURE 9. Benzoyl peroxide 0.1%, 45°C ($I = 3.26 \times 10^{-10}$).

Figure 9 shows a plot of equation (16) at 45°C , the reaction being initiated by benzoyl peroxide. Here again the points lie satisfactorily on a straight line. From these results we have

$$\left. \begin{aligned} k_3/\sqrt{k_5} &= 2.54 \times 10^4 \text{ l.}^{1/2} \text{ mol.}^{-1/2} \text{ sec.}^{-1} \text{ at } 25^\circ\text{C}, \\ k_3/\sqrt{k_5} &= 2.20 \times 10^4 \text{ l.}^{1/2} \text{ mol.}^{-1/2} \text{ sec.}^{-1} \text{ at } 45^\circ\text{C}, \end{aligned} \right\} \quad (18)$$

and

whence

$$E_3 - \frac{1}{2}E_5 = -1.3 \text{ kcal.}$$

THE PHOTOCHEMICAL AFTER-EFFECT

As stated above, the rotating sector cannot be used under conditions where the solution of oxygen is rate-determining. It is, however, possible in principle to use the photochemical after-effect to determine the velocity constants absolutely,

provided that suitable corrections can be applied. It may readily be shown from equation (2) that the photochemical after-effect ΔO_2 is given by

$$\Delta O_2 = \frac{1}{I} \left(\frac{d\{O_2\}}{dt} \right)^2 \left\{ \frac{1}{k_3[O_2]_l} + \frac{1}{k_4[TH]} \right\} \log \frac{1}{2}(1 + \xi), \quad (19)$$

in which $-d\{O_2\}/dt$ is the rate in the light, and ξ is the ratio of the rates in the light and in the dark. (If the rate in the dark is not sufficiently large, it is convenient to supply continuous weak illumination.) The total after-effect measured when the light is cut off is

$$\begin{aligned} \Delta' O_2 &= \Delta O_2 + [O_2]_{l(\text{dark})} - [O_2]_{l(\text{light})} \\ &= \Delta O_2 + s(O_2)_g - [O_2]_{l(\text{light})} \\ &= \Delta O_2 - \frac{1}{k} \frac{d\{O_2\}}{dt}, \end{aligned} \quad (20)$$

from equation (15). In deriving (19) it has been assumed that the concentration of oxygen in solution during the time of the after-effect is constant, and equal to its value in the light. This is an approximation which will be justified below.

It was found possible to measure $\Delta' O_2$ directly over the whole range of oxygen pressures. Values of ΔO_2 could be calculated from (20), since k was known from previous measurements. From (19) it follows that a plot of

$$I \Delta O_2 / \left[\left(\frac{d\{O_2\}}{dt} \right)^2 \log \frac{1}{2}(1 + \xi) \right]$$

against $1/[O_2]_l$ should be a straight line. In place of $1/[O_2]_l$ it is more convenient to plot the equivalent expression (see equation (2)), $\frac{1}{b} \left(a + 1 \left/ \frac{d\{O_2\}}{dt} \right. \right)$, where a and b have their previous significance. Figure 10 shows this plot for several series of experiments at 25° C, in which the rate of initiation was varied by a factor of about 10. All the points lie satisfactorily on the same straight line. This indicates that the change in $[O_2]_l$ during the after-effect is not serious under our conditions, for the error so introduced would be very sensitive to the rate of initiation.

From the slope and intercept of the line in figure 10, k_3 and k_4 may be found absolutely (equation (19)) and hence the other velocity constants. Values are given in table 3. The values of k_4 and k_7 agree quite satisfactorily with those obtained by the rotating sector method (equation (13)). There are a number of uncertainties in the latter method, and we therefore consider that the values in table 3 are the more reliable.

TABLE 3. VELOCITY CONSTANTS AT 25° C (l.mol.⁻¹sec.⁻¹)

k_3	6.76×10^7	k_5	7.10×10^6
k_4	13.3	k_7	2.15×10^7

The after-effects could not be measured satisfactorily at 45° C, but were always unreasonably large. The cause is not clear, but there is some evidence that photo-decomposition of tetralin hydroperoxide at 45° C produces very active catalysts

which interfere with the measurements. A similar effect probably accounts for the results of Burnett & Melville (1947*b*) on the benzoyl peroxide photosensitized polymerization of vinyl acetate. Photocatalysts are also formed in the polymerization of methyl methacrylate (Bamford & Dewar 1948*b*). It is reasonable that this effect should only occur in our experiments at the higher temperatures, since under these conditions the absorption of light by peroxide extends to larger wave-lengths.

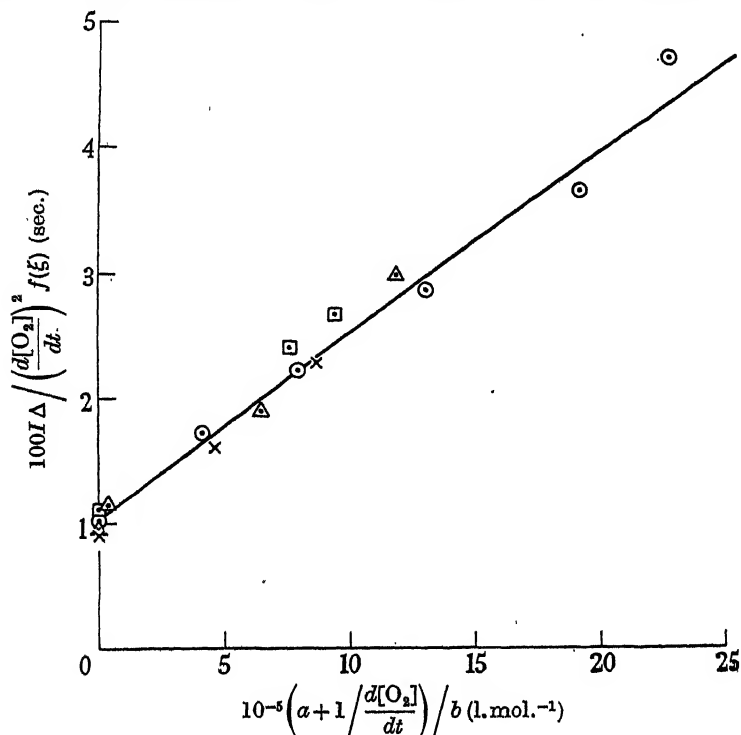


FIGURE 10. Caledon Red BN, 25° C I (mol.l.⁻¹sec.⁻¹): ○, first series 1.10×10^{-8} ; □, second series 2.08×10^{-8} ; ×, third series 3.00×10^{-8} ; △, fourth series 7.40×10^{-9} .

Accordingly, measurements of the after-effect were carried out at 0° C. Measurements were made only at high oxygen pressures, since it was considered that measurements at low oxygen pressure would not be sufficiently accurate for the calculation of the activation energies E_3 and E_5 to be significant. It will be seen later that values for these can be estimated in another way.

Equations (2) and (19) give, for high oxygen pressures,

$$\frac{\Delta O_2}{\log \frac{1}{2}(1 + \xi)} = \frac{k_4[TH]}{k_7}, \quad (21)$$

which may be used directly to obtain $E_4 - E_7$. The observed figures were

$$\left. \begin{aligned} \frac{\Delta O_2}{\log \frac{1}{2}(1 + \xi)} &= 2.49 \times 10^{-6} \text{ mol.l.}^{-1} \text{ at } 0^\circ \text{ C,} \\ \frac{\Delta O_2}{\log \frac{1}{2}(1 + \xi)} &= 4.54 \times 10^{-6} \text{ mol.l.}^{-1} \text{ at } 25^\circ \text{ C,} \end{aligned} \right\} \quad (22)$$

and

whence $E_4 - E_7 = 4.1 \text{ kcal.}$

ACTIVATION ENERGIES AND FREQUENCY FACTORS

Equations (10) and (22) give the values for E_4 and E_7 shown in table 4. Estimates of E_3 and E_5 may be made as follows. From (18), $E_5 \geq 2.6$ kcal., since E_3 cannot be negative. E_5 corresponds to the reaction between two tetralyl radicals. The activation energy for the reaction between two polystyryl radicals is 2.8 kcal. (Bamford & Dewar 1948c). In both cases the radicals are substituted benzyl radicals, and in the case of tetralin the activation energy should if anything be less, since any contribution due to diffusion would be small. Thus $E_5 \leq 2.8$ kcal. It is therefore reasonable to take $E_5 = 2.6$ and $E_3 = 0.0$ kcal.

TABLE 4. ACTIVATION ENERGIES AND FREQUENCY FACTORS

$E_3 = 0.0$ kcal.	$A_3 = 6.8 \times 10^7$
$E_4 = 4.5$ kcal.	$A_4 = 2.5 \times 10^4$
$E_5 = 2.6$ kcal.	$A_5 = 5.5 \times 10^8$
$E_7 = 0.4$ kcal.	$A_7 = 4.2 \times 10^7$

DISCUSSION

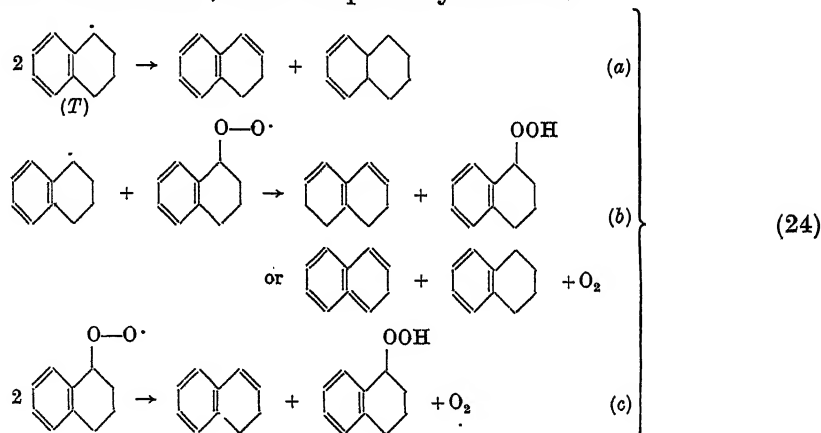
(i) Three papers by Robertson & Waters (1948*a, b, c*) on the autoxidation of tetralin have recently appeared. As their conclusions differ from ours in certain respects, some comment seems to be required.

The main difference lies in the mode of termination postulated. Robertson & Waters conclude that in the initial stage of the oxidation termination occurs by the reaction of tetralyl radicals and later by the reaction



The main argument appears to be the improbability of termination mechanisms of other types.

In the first place, equation (2) implies that reaction between tetralyl radicals can only be important at very low oxygen pressures, for under other conditions the concentration of tetralyl is extremely low. This might have been anticipated from the work of Bolland (1946). Secondly, the analytical results of Robertson & Waters not only can be interpreted in terms of our three termination reactions, but also throw some light on their mechanisms, which are probably as follows:



The formation of dihydronaphthalene in the autoxidation at high oxygen pressures and the evolution of oxygen in the decomposition of tetralin hydroperoxide (Robertson & Waters 1948*a, b*) provide strong evidence for reaction (24*c*). Robertson & Waters overlooked this possibility and regarded the presence of dihydronaphthalene as evidence for the disproportionation of tetralyl radicals. At very high rates, when the chains are short, allowance would have to be made for this evolution of oxygen. In our experiments the correction would never exceed about 3 %.

It is also evident that the decomposition of tetralin hydroperoxide (Robertson & Waters 1948*b*) can be interpreted in a similar way.

(ii) Although absolute velocity constants have now been determined for a number of radical reactions in solution (Swain & Bartlett 1946; Burnett & Melville 1947*a*; Bamford & Dewar 1947*a*, 1948*a, b, c*, and other papers in the press from this laboratory), all these have involved large radicals from polymerization processes. The measurements described in this paper seem to be the first on reactions of small radicals. It is now possible to make a preliminary survey of the field, and table 5 summarizes the data at present available for the activation energies and frequency factors of radical reactions in solution. The values for the propagation and transfer reactions in the polymerizations of styrene have been calculated from the data of Gregg & Mayo (1947), and Lewis, Walling, Cummings, Briggs & Mayo (1948) using our value (1948*a*) for the propagation constant in the polymerization. The values for vinyl acetate were obtained by Dixon-Lewis in these laboratories.

In general the frequency factors are much lower than the 'normal' value for bimolecular reactions in solution. This can hardly be a size effect, since the termination reactions which involve *two* large radicals have in fact the highest frequency factors, and since the frequency factors in the tetralin reactions are also low. Attention was drawn to these points in a previous communication (Bamford & Dewar 1949), where it was suggested that the normal frequency factor for radical reactions in solution is probably about 10^7 .

(iii) The chain lengths were in the range 25 to 600 depending on the rate of initiation. The lifetime of a TO_2 radical at 25° C is 1.02×10^{-2} sec.; this is of the same order as the life of a polystyryl radical in the polymerization of styrene at 25° C (6.2×10^{-3} sec., Bamford & Dewar 1948*a*). The lifetime of a T radical depends on the oxygen pressure. At 100 mm. and 25° C it is 2.75×10^{-5} sec. It is interesting that the life of a kinetic chain is of the order of seconds under our conditions, since it has generally been assumed that the growth of radical chains is always extremely rapid. In the thermal polymerization of styrene at 0° C the mean lifetime of the kinetic chain is 11 hr.

(iv) It is interesting to compare the rate of initiation by benzoyl peroxide with the published rate of decomposition of the peroxide. From the values quoted by Swain & Bartlett (1946) the rate of decomposition of benzoyl peroxide in 1% solution is about 10^{-9} mol.l.⁻¹sec.⁻¹. The observed rate of initiation was 2.09×10^{-10} mol.l.⁻¹sec.⁻¹. Since each molecule of peroxide could theoretically initiate two chains, there is clearly a discrepancy of a factor of 10. This might be explained in several ways (e.g. primary recombination, non-radical decomposition of the peroxide, or induced decomposition of the peroxide by radicals), but in any case it throws doubt on the

TABLE 5

reaction	E_1 kcal.	$\log_{10} A$
polymerization of styrene:		
propagation	6.5	6.01
transfer	14.2	7.18
termination	2.8	8.49
polymerization of vinyl acetate:		
propagation	3.0	5.84
transfer	5.9	3.86
termination	0.0	8.30
co-polymerization, reactions of polystyryl radicals with:		
methyl methacrylate	6.0	6.01
methyl acrylate	6.1	5.90
diethyl maleate	7.2	5.60
diethyl fumarate	5.4	5.83
<i>p</i> -chlorostyrene	6.1	5.91
chain transfer reactions of styrene with solvents:		
benzene	21.3	15.07
toluene	16.6	10.04
ethylbenzene	12.0	4.74
isopropylbenzene	12.0	4.92
tertbutylbenzene	20.2	14.68
diphenylmethane	10.2	3.25
triphenylmethane	11.6	6.25
fluorene	9.6	5.82
cyclohexane	19.9	13.22
<i>n</i> -heptane	11.5	3.57
ethylene dichloride	21.2	11.10*
ethylene dibromide	16.2	8.48*
carbon tetrachloride	10.7	6.88*
carbon tetrabromide	6.7	6.59*
autoxidation of tetralin:		
k_3	0	7.82
k_4	4.5	4.40
k_5	2.6	8.74
k_7	0.4	7.62

* Bamford & Dewar (1947b).

validity of work in which the rate of chains starting has been equated to the rate of decomposition of the peroxide (cf. Swain & Bartlett 1946; Bolland 1948. The work of the former authors will be discussed in this connexion elsewhere).

(v) The techniques described in this paper can clearly be extended to other free radical reactions, e.g. the addition of halogen compounds to olefines, and work along such lines is in progress here.

REFERENCES

- Bamford, C. H. & Dewar, M. J. S. 1947*a* *Disc. Faraday Soc.* 2, 310.
 Bamford, C. H. & Dewar, M. J. S. 1947*b* *Disc. Faraday Soc.* 2, 314.
 Bamford, C. H. & Dewar, M. J. S. 1948*a* *Proc. Roy. Soc. A*, 192, 309.
 Bamford, C. H. & Dewar, M. J. S. 1948*b* *Proc. Roy. Soc. A*, 192, 329.
 Bamford, C. H. & Dewar, M. J. S. 1948*c* In course of publication.

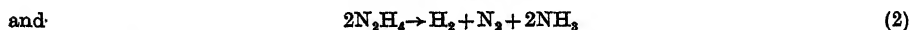
- Bamford, C. H. & Dewar, M. J. S. 1949 *Nature*, **163**, 256.
 Bolland, J. L. 1946 *Proc. Roy. Soc. A*, **186**, 218.
 Bolland, J. L. 1948 *Trans. Faraday Soc.* **44**, 669.
 Bolland, J. L. & Gee, G. 1946 *Trans. Faraday Soc.* **42**, 236, 244.
 Bolland, J. L. & ten Have, P. 1947a *Trans. Faraday Soc.* **43**, 201.
 Bolland, J. L. & ten Have, P. 1947b *Disc. Faraday Soc.* **2**, 252.
 Briers, F., Chapman, D. L. & Walters, E. 1926 *J. Chem. Soc.* p. 562.
 Burnett, G. M. & Melville, H. W. 1947a *Proc. Roy. Soc. A*, **189**, 456.
 Burnett, G. M. & Melville, H. W. 1947b *Proc. Roy. Soc. A*, **189**, 474.
 Fischer, F. & Pfeleiderer, G. 1922 *Z. anorg. Chem.* **124**, 61.
 George, P. 1946a *Proc. Roy. Soc. A*, **185**, 337.
 George, P. 1946b *Trans. Faraday Soc.* **42**, 210.
 George, P., Rideal, E. K. & Robertson, A. 1946 *Proc. Roy. Soc. A*, **185**, 288.
 George, P. & Robertson, A. 1946a *Proc. Roy. Soc. A*, **185**, 309.
 George, P. & Robertson, A. 1946b *Trans. Faraday Soc.* **42**, 217.
 Gregg, R. A. & Mayo, F. R. 1947 *Disc. Faraday Soc.* **2**, 328.
 Lewis, F. M., Walling, C., Cummings, W., Briggs, E. R. & Mayo, F. R. 1948 *J. Amer. Chem. Soc.* **70**, 1519.
 Noyes, W. A. & Leighton, P. A. 1941 *Photochemistry of gases*. New York: Reinhold.
 Robertson, A. & Waters, W. A. 1946 *Trans. Faraday Soc.* **42**, 201.
 Robertson, A. & Waters, W. A. 1948a *J. Chem. Soc.* p. 1574.
 Robertson, A. & Waters, W. A. 1948b *J. Chem. Soc.* p. 1578.
 Robertson, A. & Waters, W. A. 1948c *J. Chem. Soc.* p. 1585.
 Swain, C. G. & Bartlett, P. D. 1946 *J. Amer. Chem. Soc.* **68**, 2381.
 Weber, G. 1948 *Trans. Faraday Soc.* **44**, 185.

The dissociation energy of the N-N bond in hydrazine

BY M. SZWARC, *Chemistry Department, University of Manchester*

(Communicated by M. G. Evans, F.R.S.—Received 19 January 1949)

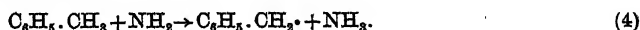
The thermal decomposition of hydrazine was investigated over a temperature range of 630 to 780° C and at pressures of a few mm. Hg. The experiments were carried out in a flow system, toluene being used as a carrier gas. This technique makes it possible to discriminate between the heterogeneous decomposition of hydrazine



and the homogeneous decomposition to



The NH_2 radicals produced by the latter process were removed, in the presence of excess toluene, by the rapid reaction (4)



Thus the rate of formation of dibenzyl measures the rate of reaction (3). The study of the stoichiometry of the overall process, of the kinetics of various steps and of the effect of packing the reaction vessel led to the conclusion that (3) is a homogeneous, unimolecular gas reaction, the rate constant being $4 \times 10^{12} \exp(60,000/RT)$. Assuming that the recombination of NH_2 radicals does not involve any energy of activation, it is found that $D(\text{NH}_2\text{-NH}_2) = 60 \pm 3$ kcal./mole. This value in conjunction with the relevant thermochemical data leads to the heat of formation of NH_2 radical as 41 kcal./mole and to $D(\text{NH}_2\text{-H}) = 104 \pm 2$ kcal./mole. The latter is good evidence in support of Gaydon's value of 225 kcal./mole for the heat of dissociation of N_2 .

INTRODUCTION

Very few data have been published on the subject of the thermal decomposition of hydrazine. Elgin & Taylor (1929) first observed the decomposition of this substance in a silica vessel at 250° C, and they found N_2 and NH_3 to be the main products. From the relative quantities of N_2 and NH_3 they deduced that the overall process is represented by the equation



However, they also reported the formation of small quantities of H_2 (about 5 %), which indicated that the main process, represented by equation (1), was accompanied by a much slower decomposition corresponding to equation (2)



Subsequent work carried out in Taylor's laboratory by Askey (1930) confirmed the overall stoichiometry of equation (1) and revealed that the decomposition of hydrazine in a silica bulb at 300° C is a heterogeneous process, which takes place on the walls of the reaction vessel and obeys first-order kinetics. The same worker found also that the thermal decomposition of hydrazine on a platinum or tungsten wire (at 200 and 380° C respectively) produced N_2 , H_2 and NH_3 according to the overall equation



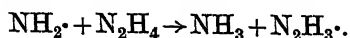
Gedye & Allibone (1931) used the thermal decomposition of hydrazine (in a glass vessel at 400° C) as an analytical method for determining the quantities of hydrazine in a gas mixture. They also adopted the equation $3N_2H_4 = N_2 + 4NH_3$. Further support for this scheme was provided by Gedye & Rideal (1932), although in one single experiment, in which the reaction vessel had been evacuated and sparked for some time with a Tesla coil, they observed that the decomposition proceeded according to equation (2).

More recently, in a paper published by Birse & Melville (1940), data are presented which enable one to calculate the rate constant of the thermal decomposition of hydrazine in a silica vessel at 75, 100 and 218° C. From these data we estimate the energy of activation of this process as about 18 kcal./mole.

The present investigation was concerned with the thermal decomposition of hydrazine in a temperature range of 620 to 770° C, i.e. much higher than those used in the previous studies. The essential new feature of the method chosen in this investigation lies in the use of toluene as a carrier gas. If hydrazine decomposes into two NH_2 radicals

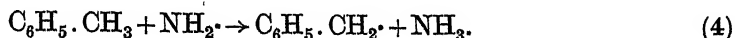


then, with hydrazine alone or in the presence of a neutral carrier gas, the NH_2 radicals could react with undecomposed hydrazine yielding, most probably, NH_3 and N_2H_3 radicals



The subsequent behaviour of the N_2H_3 radicals may lead to a chain process, and at any rate the interpretation of the results obtained under such conditions would be

complicated and probably ambiguous. In the presence of toluene as a carrier gas, however, it might be expected that NH_2 radicals would be removed by the following reaction:



The stable benzyl radicals formed in this reaction would eventually dimerize, as was shown in previous studies (Horrex & Szwarc unpublished; Szwarc 1947, 1948, 1949); and the mechanism of the decomposition would thus be greatly simplified (see, for example, Szwarc 1949). This expectation is borne out by the results reported in this paper.

EXPERIMENTAL

The experimental technique was essentially the same as that already described by the present writer in a paper dealing with the thermal decomposition of ethylbenzene (Szwarc 1949), and accordingly only a brief account of the method need be given here.

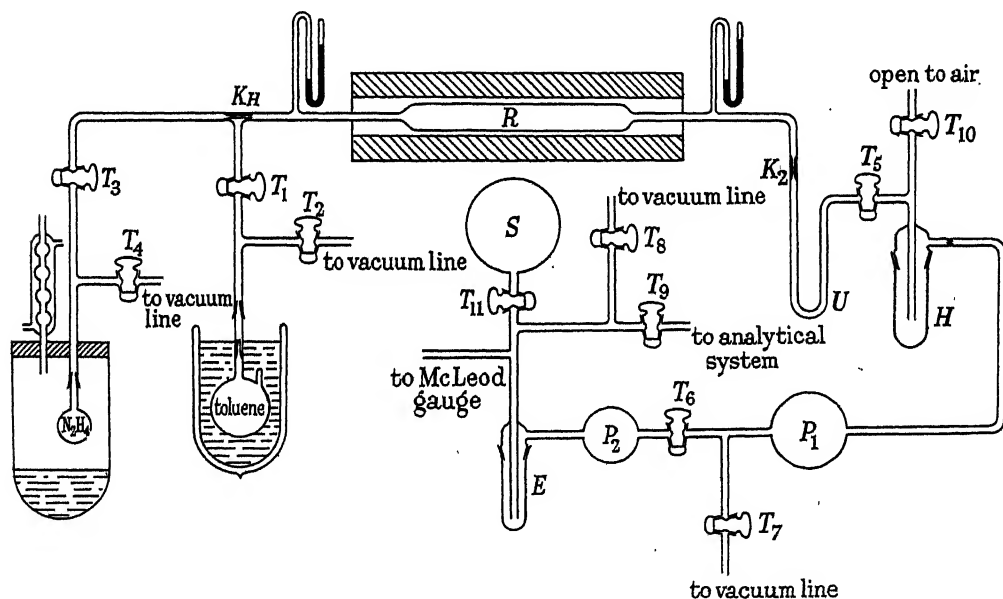


FIGURE 1

The apparatus is shown diagrammatically in figure 1. The vapour of anhydrous hydrazine was introduced through a fine capillary K_H into a large excess of toluene vapour. The mixture of both compounds, maintained under a pressure of a few mm. Hg was made to flow continuously through a silica tube R , heated by means of a cylindrical electric furnace. The gases leaving the reaction vessel were lead through a glass tube and glass capillary K_2 (both heated by a nichrome winding to about 80°C) into a small trap U cooled by ice. The non-volatile products (e.g. dibenzyl) were deposited in this trap, while the undecomposed toluene and hydrazine passed through it and were condensed in a large trap H cooled to about -80°C by a mixture of acetone and solid CO_2 .

The gases produced in the decomposition (H_2 , N_2 and NH_3) were continuously removed by a system of two mercury diffusion pumps P_1 and P_2 and pumped through a trap E cooled with liquid air into a storage vessel S . Ammonia was frozen out in the trap E , and thus the amount of $\text{H}_2 + \text{N}_2$ formed in the pyrolysis was estimated by measuring the pressure at the end of the run.* Thereafter part of the $\text{H}_2 + \text{N}_2$ mixture was admitted into an analytical device in which the percentage of H_2 could be estimated; the remaining gas being pumped off. The liquid air surrounding trap E was removed and replaced by an ice-water bath. The ammonia condensed in E evaporated and the quantity obtained was estimated by measuring the pressure in a known volume. The gas was shown to consist entirely of ammonia by absorbing it in dilute sulphuric acid.

The trap H contained all the undecomposed hydrazine, toluene, and small quantities of dibenzyl, which passed through trap U . In order to estimate the amount of hydrazine the contents of the trap H were extracted with dilute sulphuric acid and the aqueous layer was analyzed for hydrazine by titration with HIO_3 in acid solution (Bray & Cuy 1924). To determine the quantity of dibenzyl the toluene was removed by distillation *in vacuo* at ice temperature and the residue was weighed. The weight of this residue was added to the weight of the bulk of the dibenzyl given by the increase in the weight of trap U . The accuracy of this determination was about 1 to 2 mg. The solid in trap U was identified as dibenzyl by its melting point, the unpurified material melting at about 50°C (m.p. of dibenzyl 51°C).

The material used

Anhydrous hydrazine was prepared from a commercial 60 % solution of hydrazine hydrate according to the method described by Wenner & Beckman (1932). Analysis of the final product showed it to contain 98 to 99 % of anhydrous hydrazine.

Toluene was prepared as in previous studies by the repeated pyrolysis of 'sulphur free' toluene at 810 to 820°C followed by careful distillation of the pyrolyzed material. Blank experiments showed that the thermal decomposition of toluene could be neglected when compared with the decomposition of hydrazine at the temperatures used in this investigation.

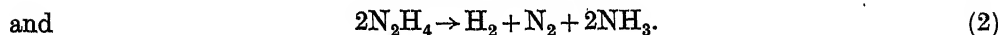
THE RESULTS AND THEIR INTERPRETATION

The decomposition of hydrazine in these experiments produced the following products: H_2 , N_2 , NH_3 and dibenzyl. In the experiments carried out at higher temperatures the formation of small quantities of CH_4 was also observed. Table 1 illustrates the results, giving the percentage of H_2 in the $\text{H}_2 + \text{N}_2$ mixture, the molar ratio of NH_3 to $\text{H}_2 + \text{N}_2$, and the molar ratio of dibenzyl to NH_3 . The following points should be noted:

1. The formation of H_2 and N_2 suggests that even at the high temperatures used in the present research there was a considerable amount of heterogeneous decomposition of hydrazine on the walls of the reaction vessel.

* The pumping out of the trap H was continued for half an hour, after the flow of toluene and hydrazine had been interrupted. The tap T_4 was then closed and the pressure of the collected gas measured.

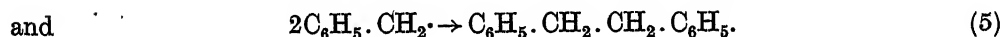
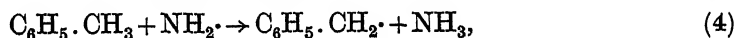
2. It seems that both modes of heterogeneous decomposition mentioned in the introduction participate in the actual process:



The occurrence of reaction (2) is shown by the appearance of H_2 amongst the products of decomposition. The fact that the amount of N_2 always exceeded that of H_2 is regarded as evidence for the participation of reaction (1).

3. The observed increase in the percentage of H_2 as the temperature of decomposition was increased indicates that the reaction (2) has a higher energy of activation than reaction (1). This explains why previous investigators, who worked at much lower temperatures, observed mainly reaction (1).

4. The marked increase, with increasing temperature, of the molar fraction of dibenzyl in the decomposition products indicates that the process by which it is formed has a high energy of activation. We assume that dibenzyl is produced as a result of the following three reactions:



This assumption is crucial for our further conclusions, and its justification will be fully discussed later.

If we assume that the decomposition of hydrazine takes place according to equations (1), (2) and (3), the latter process being followed, in the presence of an

TABLE I

run	T ($^{\circ}\text{K}$)	% H_2	$\text{NH}_3/(\text{H}_2 + \text{N}_2)$	dibenzyl/ NH_3
56	894	14	3.06	0.0072
55	896	16	3.34	0.0185
15	900	25	2.56	0.031
17-18	903	22	2.41	0.023
20	951	32	2.14	0.096
19	955	32	2.28	0.131
22	955	32	2.30	0.097
27	995	40	2.03	0.127
53	1004	32	2.57	0.127
38	1007	31	3.06	0.20
51	1008	35	2.98	0.15
26	1008	38	2.40	0.17
31	1009	35	2.62	0.155
30	1010	35	2.84	0.16
5-6	1044	40	2.24	0.25
7-8	1047	44	2.10	0.26
59	1050	35	2.98	0.24
57	1053	40	2.84	0.25
3-4	1057	42	2.56	0.25

This table contains only a few of the thirty-two results obtained in the region of 1000 to 1010 $^{\circ}$ K.

excess of toluene, by reactions (4) and (5), we are able to calculate the amounts of NH_3 produced and of hydrazine decomposed from a knowledge of the quantities of H_2 , N_2 and dibenzyl formed in the pyrolysis. In this scheme all the hydrogen is assumed to be produced by reaction (2). The difference between the observed amounts of N_2 and H_2 is taken as a measure of the quantity of N_2 produced by reaction (1). The amount of dibenzyl formed should be the same as the quantity of hydrazine decomposed according to reaction (3) and it should be also equal to half the number of moles of NH_3 formed by reaction (4). This method of interpretation is illustrated by calculations from the results of two experiments, one chosen from the lowest and the second from the highest temperature region.

Experiment 55. $T = 896^\circ \text{K}$

observed amount in millimoles		equation of decomposition	NH_3 calculated in millimoles	N_2H_4 decomposed in millimoles
H_2	0.13	$2\text{N}_2\text{H}_4 = \text{H}_2 + \text{N}_2 + 2\text{NH}_3$	$2 \times 0.13 = 0.26$	$2 \times 0.13 = 0.26$
N_2	$0.68 = 0.13 + 0.55$	$3\text{N}_2\text{H}_4 = \text{N}_2 + 4\text{NH}_3$	$4 \times 0.55 = 2.20$	$3 \times 0.55 = 1.65$
dibenzyl	0.05	$\begin{cases} \text{N}_2\text{H}_4 = 2\text{NH}_2; \\ 2\text{C}_6\text{H}_5 \cdot \text{CH}_3 + 2\text{NH}_2 \\ = 2\text{NH}_3 + \text{dibenzyl} \end{cases}$	$2 \times 0.05 = 0.10$	$1 \times 0.05 = 0.05$
			total = 2.56	total = 1.96
			observed = 2.70	

Experiment 59. $T = 1050^\circ \text{K}$

observed amount in millimoles		equation of decomposition	NH_3 calculated in millimoles	N_2H_4 decomposed in millimoles
H_2	0.25	$2\text{N}_2\text{H}_4 = \text{H}_2 + \text{N}_2 + 2\text{NH}_3$	$2 \times 0.25 = 0.50$	$2 \times 0.25 = 0.50$
N_2	$0.43 = 0.25 + 0.18$	$3\text{N}_2\text{H}_4 = \text{N}_2 + 4\text{NH}_3$	$4 \times 0.18 = 0.72$	$3 \times 0.18 = 0.54$
CH_4	0.04	—	—	—
dibenzyl	0.48	$\begin{cases} \text{N}_2\text{H}_4 = 2\text{NH}_2 \\ 2\text{C}_6\text{H}_5 \cdot \text{CH}_3 + 2\text{NH}_2 \\ = 2\text{NH}_3 + \text{dibenzyl} \end{cases}$	$2 \times 0.48 = 0.96$	$1 \times 0.48 = 0.48$
			total = 2.18	total = 1.52
			observed = 2.02	

Table 2 contains the observed quantities of H_2 , N_2 , CH_4 , dibenzyl and NH_3 and the calculated quantities of NH_3 . The agreement between the observed and calculated quantities of NH_3 , given in the last two columns of table 2, provides a strong argument in favour of our assumptions. Further evidence favouring our interpretation is provided in tables 3 and 4. The third column of table 3 contains the quantities of hydrazine decomposed calculated by the method explained above. Column 4 of this table gives the quantities of undecomposed hydrazine, recovered from the trap H , and determined by direct titration. The last column contains the quantities of hydrazine introduced (referred to later as Q_c) equal to the sum of decomposed and recovered hydrazine. Owing to technical difficulties it was not feasible to estimate the quantity of hydrazine introduced in each individual run. It was possible, however, to estimate the total amount for a set of about 4 to 15 runs by direct weighing of the hydrazine reservoir. Table 4 contains data showing the total amount of hydrazine introduced in each of several sets of runs. These quantities are compared with the ΣQ_c obtained by summing up the corresponding values from the last column of table 3. The agreement seems to be satisfactory.

TABLE 2

The quantities are given in millimoles

run	T ($^{\circ}\text{K}$)	H_2	N_2	CH_4	dibenzyl	NH_3	
						obs.	calc.
56	894	0.10	0.63	—	0.016	2.23	2.35
55	896	0.13	0.68	—	0.05	2.70	2.56
15	900	0.22	0.66	—	0.07	2.25	2.34
17-18	903	0.22	0.78	—	0.055	2.41	2.79
20	951	0.29	0.62	—	0.187	1.95	2.27
19	955	0.29	0.61	—	0.27	2.06	2.40
22	955	0.265	0.56	—	0.185	1.90	2.08
27	995	0.26	0.36	0.02	0.16	1.26	1.14
53	1004	0.16	0.33	0.01	0.16	1.26	1.32
52	1006	0.15	0.33	0.01	0.20	1.30	1.42
38	1007	0.16	0.33	0.02	0.30	1.50	1.60
50	1007	0.15	0.33	0.01	0.17	1.25	1.36
54	1006	0.14	0.42	0.01	0.23	1.77	1.86
51	1008	0.18	0.32	0.01	0.22	1.49	1.36
45	1000	0.18	0.30	0.01	0.21	1.20	1.26
37	1004	0.15	0.29	0.01	0.20	1.25	1.26
47	999	0.14	0.35	0.01	0.13	1.40	1.34
26	1008	0.68	1.06	0.05	0.71	4.19	4.30
31	1009	0.15	0.27	0.01	0.17	1.10	1.12
32	1002	0.37	0.58	0.02	0.40	2.28	2.38
30	1010	0.16	0.28	0.01	0.20	1.25	1.20
49	1006	0.20	0.35	0.01	0.20	1.21	1.40
25	1010	0.85	1.25	0.04	0.63	4.50	4.56
24	1008	0.52	0.79	0.04	0.51	2.84	3.14
33	1010	0.24	0.38	0.02	0.16	1.35	1.36
36	1002	0.26	0.42	0.025	0.17	1.63	1.50
28	1007	0.32	0.47	0.02	0.29	1.50	1.82
35	1006	0.30	0.50	0.02	0.14	1.68	1.68
29	1011	0.515	0.745	0.025	0.42	2.76	2.79
48	1008	0.24	0.40	0.01	0.39	1.78	1.90
46	1003	0.19	0.33	0.01	0.34	1.58	1.62
40	1007	0.21	0.34	0.01	0.325	1.49	1.59
39	1006	0.21	0.36	—	0.38	1.62	1.78
60	1007	0.24	0.43	—	0.09	1.40	1.42
61	1009	0.21	0.41	—	0.09	1.45	1.40
62	1009	0.14	0.49	—	0.18	1.60	2.04
63	1015	0.19	0.41	—	0.12	1.30	1.50
64	1007	0.16	0.45	—	0.13	1.45	1.74
65	1010	0.29	0.69	—	0.18	2.70	2.54
66	1007	0.20	0.57	—	0.12	2.05	2.12
5-6	1044	0.41	0.58	0.03	0.55	2.22	2.60
7-8	1047	2.26	2.70	0.20	2.73	10.4	11.8
59	1050	0.25	0.43	0.04	0.48	2.02	2.18
57	1053	0.30	0.42	0.04	0.52	2.04	2.12
3-4	1057	0.46	0.56	0.09	0.64	2.60	2.22

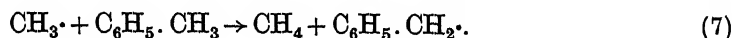
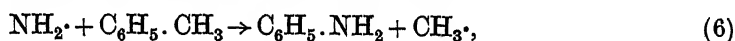
TABLE 3

run	T ($^{\circ}\text{K}$)	N_2H_4 decomp. (millimoles)	N_2H_4 recovered (undecomp.) (millimoles)	$Q_0 \text{N}_2\text{H}_4$ introduced (millimoles)
56	894	1.81	5.80	7.61
55	896	1.96	8.10	10.06
15	900	1.87	9.31	11.18
17-18	903	2.17	10.44	12.61
20	951	1.76	4.75	6.51
19	955	1.82	6.50	8.32
22	955	1.59	6.60	8.19
27	995	0.98	1.91	2.89
53	1004	0.99	1.04	2.03
52	1006	1.04	1.34	2.38
38	1007	1.13	1.52	2.65
50	1007	1.01	0.87	1.88
54	1006	1.35	1.20	2.55
51	1008	1.00	1.15	2.15
45	1000	0.93	1.15	2.08
37	1004	0.92	1.96	2.88
47	999	1.01	1.37	2.38
26	1008	3.21	5.50	8.71
31	1009	0.83	0.86	1.69
32	1002	1.77	3.62	5.39
30	1010	0.88	1.15	2.03
49	1006	1.05	1.52	2.57
25	1010	3.53	4.80	8.33
24	1008	2.36	4.00	6.36
33	1010	1.06	1.50	2.56
36	1002	1.17	1.74	2.91
28	1007	1.38	2.18	3.56
35	1006	1.34	1.10	2.44
29	1011	2.14	2.78	4.92
48	1008	1.35	1.80	3.15
46	1003	1.14	1.80	2.94
40	1007	1.14	0.18	1.32
39	1006	1.25	0.24	1.49
60	1007	1.14	0.68	1.82
61	1009	1.11	0.15	1.26
62	1009	1.51	0.59	2.10
63	1015	1.16	0.22	1.38
64	1007	1.32	0.40	1.72
65	1010	1.96	0.42	2.38
66	1007	1.69	0.29	1.88
5-6	1044	1.88	1.11	2.99
7-8	1047	8.57	4.90	13.47
59	1050	1.52	0.57	2.09
57	1053	1.46	0.61	2.07
3-4	1057	1.82	0.69	2.51

TABLE 4

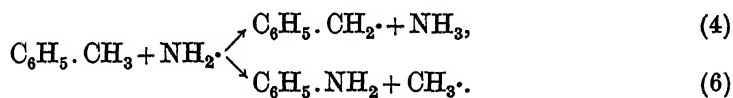
runs	N ₂ H ₄ introduced exper. (millimoles)	N ₂ H ₄ introduced theor. ΣQ_0 (millimoles)
17-20	29.2	27.4
27-40	40.0	38.9
45-57	46.5	43.9
60-66	12.0	12.5

The appearance of CH₄ merits some further discussion. Previous work, as well as the blank tests performed during the present work, proved definitely that the observed quantities of CH₄ could not be attributed to the thermal decomposition of toluene. It could be suggested that CH₄ is produced by the following reactions:



To test this suggestion the toluene, recovered from high temperature experiments, was extracted first with water, in order to remove hydrazine, and then with dilute hydrochloric acid. The latter solution was then diazotized and coupled with β naphthol in the usual way. The appearance of the red colour indicated the presence of aniline in the original condensate. It was also shown that this colour was not developed if a mixture of toluene and hydrazine that had not previously been pyrolyzed was used.

However, judging from the intensity of the red colour the amount of aniline present was only about 1/10 of that which could be expected on the basis of the quantity of CH₄ produced. Taking the amount of aniline determined experimentally as a measure of occurrence of reaction (6) the relative probabilities of reactions (4) and (6) were estimated, from the quantities of dibenzyl and aniline produced, to be at least 200:1



Assuming that reactions (4) and (6) have equal steric factors one is able to calculate on the basis of the above relative probabilities the difference in the activation energies $E_4 - E_6$. The obtained value of about 10 kcal./mole is obviously too high for reactions of this type. It is likely, therefore, that these probabilities are determined by steric factors and not by the differences in the activation energy of the two processes.

THE KINETICS OF THE DECOMPOSITION OF HYDRAZINE

It was stated by Askey (1930), that the heterogeneous decomposition of hydrazine obeys first-order kinetics and we assume that this is true also in our experiments. Furthermore, we consider that the rate of formation of dibenzyl, in the presence of a large excess of toluene, is governed by the rate of the reaction



which is also assumed to be a first-order process. The successful interpretation of the experimental data on the basis of these assumptions will be considered as evidence for their validity.

Using this kinetic scheme we are able to calculate the composite rate constant of the overall decomposition, K_{total} , and the rate constant of the hydrazine decomposition proceeding according to equation (3), denoted by $K_{\text{N}_2\text{H}_4}$, using the following formulae:

$$K_{\text{total}} = \frac{1}{t} \ln \frac{\text{amount of N}_2\text{H}_4 \text{ introduced}}{\text{amount of N}_2\text{H}_4 \text{ undecomposed}},$$

$$K_{\text{total}} = \frac{1}{(\text{N}_2\text{H}_4)} \frac{\Delta(\text{total N}_2\text{H}_4 \text{ decomposed})}{\Delta t}$$

$$= \frac{1}{(\text{N}_2\text{H}_4)} \left\{ \frac{\Delta(\text{N}_2\text{H}_4 \text{ decomposed homogen.})}{\Delta t} + \frac{\Delta(\text{N}_2\text{H}_4 \text{ decomposed heterogen.})}{\Delta t} \right\},$$

therefore

$$K_{\text{N}_2\text{H}_4} = K_{\text{total}} \frac{\text{amount of N}_2\text{H}_4 \text{ decomposed according to equation (3)}}{\text{total amount of N}_2\text{H}_4 \text{ decomposed}}$$

$$= K_{\text{total}} \frac{\text{amount of dibenzyl formed}}{\text{total amount of N}_2\text{H}_4 \text{ decomposed}},$$

t denotes the time of contact and is calculated from the measured rates of evaporation of toluene and of hydrazine, and the pressure and temperature in the reaction vessel. The quantity $\frac{\text{amount of dibenzyl formed}}{\text{total amount of N}_2\text{H}_4 \text{ decomposed}}$ will henceforth be described as the 'fraction of homogeneous decomposition'. The necessary data: amounts of N_2H_4 introduced, of undecomposed (recovered) N_2H_4 and total amounts of N_2H_4 decomposed, are tabulated in table 3; the amounts of dibenzyl formed are given in table 2. Table 5 shows the values of K_{total} calculated in this way, the 'fraction of homogeneous decomposition' and $K_{\text{N}_2\text{H}_4}$ for experiments carried out at temperatures near 1008°K . The last column of this table contains the values of $K_{\text{N}_2\text{H}_4}$ re-calculated for an arbitrarily chosen temperature of 1008°K . This re-calculation involved a short extrapolation which was carried out on the basis of 60 kcal./mole as an activation energy for the process of formation of dibenzyl; a value obtained by plotting $\log K_{\text{N}_2\text{H}_4}$ against $1/T$ (see table 6 and figure 2). The following conclusions may be drawn from table 5:

1. Although the partial pressure of hydrazine varied from 0.05 to 0.78 mm. Hg, the pressure of toluene varied from 5 up to 15 mm. Hg, and the time of contact varied from 0.26 up to 1.1 sec., there does not appear to be any trend in the values of K_{total} . These results justify our assumption concerning the first-order kinetics of the overall process.

2. The packing of the reaction vessel (see runs 60 to 66), which increased the surface about $2\frac{1}{2}$ times, caused a roughly proportional increase in K_{total} . The average K_{total} for twenty-three experiments carried out in the unpacked reaction vessel is 2.1 sec.^{-1} ; while the average for seven experiments in the packed reaction

TABLE 5

run	T ($^{\circ}$ K)	P_{anvase} (mm. Hg)	P_{NH_4} (mm. Hg)	time of contact (sec.)	K_{total} (sec. $^{-1}$)	fraction of homogeneous decomp.	K_{NH_4} (sec. $^{-1}$)	K_{NH_4} for 1008 $^{\circ}$ K (sec. $^{-1}$)
toluene pressure decreased	53	1004	4.9	0.12	2.2	0.170	0.38	0.43
	52	1006	4.9	0.14	1.9	0.192	0.37	0.39
	38	1007	7.0	0.045	2.0	0.256	0.51	0.53
	50	1007	7.2	0.05	2.8	0.168	0.48	0.49
	54	1006	7.2	0.09	2.7	0.170	0.46	0.49
	51	1008	7.2	0.16	2.3	0.220	0.51	0.51
	45	1000	7.3	0.17	1.9	0.226	0.44	0.56
	37	1004	7.2	0.20	1.4	0.220	0.31	0.35
	47	999	7.3	0.23	2.1	0.129	0.27	0.35
	26	1008	7.0	0.23	1.7	0.220	0.38	0.38
N_2H_4 partial pressure increasing	31	1009	7.3	0.23	2.5	0.204	0.50	0.49
	32	1002	7.1	0.26	1.5	0.226	0.33	0.40
	30	1010	7.3	0.26	2.1	0.224	0.42	0.40
	49	1006	7.2	0.27	1.9	0.190	0.37	0.39
	25	1010	7.0	0.29	2.1	0.179	0.38	0.36
	24	1008	7.3	0.34	1.7	0.216	0.37	0.37
	33	1010	6.5	0.36	2.4	0.151	0.37	0.35
	36	1002	7.6	0.51	1.8	0.145	0.27	0.34
	28	1007	7.4	0.59	1.8	0.210	0.38	0.40
	35	1006	7.5	0.75	2.8	0.104	0.30	0.32
toluene pressure increased	29	1011	7.2	0.78	2.1	0.200	0.42	0.38
	48	1008	11.3	0.45	1.9	0.288	0.55	0.55
	46	1003	11.9	0.40	1.7	0.298	0.50	0.58
	40	1007	15.5	0.43	1.8	0.232	0.43	0.44
	39	1006	15.0	0.45	1.4	0.304	0.44	0.47
	60	1007	6.5	0.10	4.0	0.079	0.32	0.33
	61	1009	8.5	0.12	6.0	0.080	0.48	0.47
	62	1009	8.2	0.15	4.0	0.119	0.48	0.47
	63	1015	7.8	0.10	5.2	0.103	0.54	0.47
	64	1007	7.7	0.16	4.7	0.098	0.47	0.48
reaction vessel packed. Surface about $2\frac{1}{2}$ times increased	65	1010	7.5	0.20	6.0	0.092	0.55	0.52
	66	1007	8.6	0.23	5.8	0.071	0.41	0.43

vessel is 5.1 sec.⁻¹. This is a direct proof that processes (1) and (2) are indeed heterogeneous.

3. The values for $K_{N_2H_4}$ re-calculated for 1008° K are even more reproducible than K_{total} . They do not show any trend either with the change of partial pressure of hydrazine (varied by a factor of 18), or with the change in time of contact. This justifies our second assumption, namely the first-order kinetics of the process of formation of dibenzyl. There is a slight increase in $K_{N_2H_4}$ with the increase in toluene pressure (compare runs 53 and 52 with 48 and 46), but these variations are not of sufficient magnitude to be considered important.

4. The packing of the reaction vessel seems to be without any influence on the $K_{N_2H_4}$. This observation strongly suggests that the rate determining process in the formation of dibenzyl is a homogeneous gas reaction. It is interesting to note that the random variations in K_{total} are much greater than the variations in $K_{N_2H_4}$. This seems to be an indirect indication of the heterogeneous nature of processes (1) and (2) and of the homogeneous nature of reaction (3).

TABLE 6

run	<i>T</i> (° K)	<i>P</i> _{toluene} (mm. Hg)	<i>P</i> _{N₂H₄} (mm. Hg)	time of contact (sec.)	<i>K</i> _{total} (sec. ⁻¹)	fraction of homo- geneous decomp.	<i>K</i> _{N₂H₄} (sec. ⁻¹)	<i>E</i> (kcal./mole)
56	894	7.0	0.10	0.33	0.82	0.009	0.007	61.1
55	896	7.0	0.16	0.35	0.61	0.025	0.015	59.9
15	900	7.0	0.15	0.30	0.61	0.038	0.023	59.4
17-18	903	7.0	0.11	0.30	0.61	0.024	0.015	60.3
20	951	7.0	0.44	0.29	1.10	0.106	0.12	59.6
19	955	6.8	0.16	0.28	0.89	0.148	0.13	59.7
22	955	7.0	0.42	0.29	0.74	0.116	0.09	60.4
27	995	7.1	0.46	0.27	1.6	0.163	0.25	60.9
average at 1008		~ 7.0	—	~ 0.27	2.1	0.20	0.41	60.7
5-6	1044	7.1	0.19	0.27	3.5	0.29	1.0	61.0
7-8	1047	7.1	0.16	0.27	3.8	0.32	1.3	60.6
59	1050	7.0	0.10	0.27	4.8	0.32	1.6	60.4
57	1053	7.1	0.07	0.26	4.7	0.34	1.6	60.5
3-4	1057	7.0	0.19	0.27	4.8	0.33	1.6	60.8

Having presented evidence in support of our assumptions concerning the kinetics of the reaction, we can now discuss the problem of activation energy. Table 6 contains the values of K_{total} , 'fraction of homogeneous decomposition' and $K_{N_2H_4}$ for four temperature regions from 900 to 1050° K. There is a small increase in K_{total} , but a very marked increase in the 'fraction of homogeneous decomposition' and in $K_{N_2H_4}$. The last column of table 6 contains the values of activation energy corresponding to $K_{N_2H_4}$ calculated on the assumption that the frequency factor is 5×10^{12} sec.⁻¹. This particular value of the frequency factor was chosen in consideration of the results obtained in a previous investigation (Szwarc 1949). The plot of $\log K_{N_2H_4}$ against $1/T$ is given in figure 2. The best straight line corresponds to an activation energy of 60 ± 3 kcal./mole and a frequency factor of 4×10^{12} sec.⁻¹.

The agreement between the experimental frequency factor obtained from figure 2 and that which is expected for a unimolecular process on theoretical grounds (10^{12} to 10^{13} sec. $^{-1}$) confirms that $K_{N_2H_4}$ represents the rate constant of the homogeneous, unimolecular gas reaction.

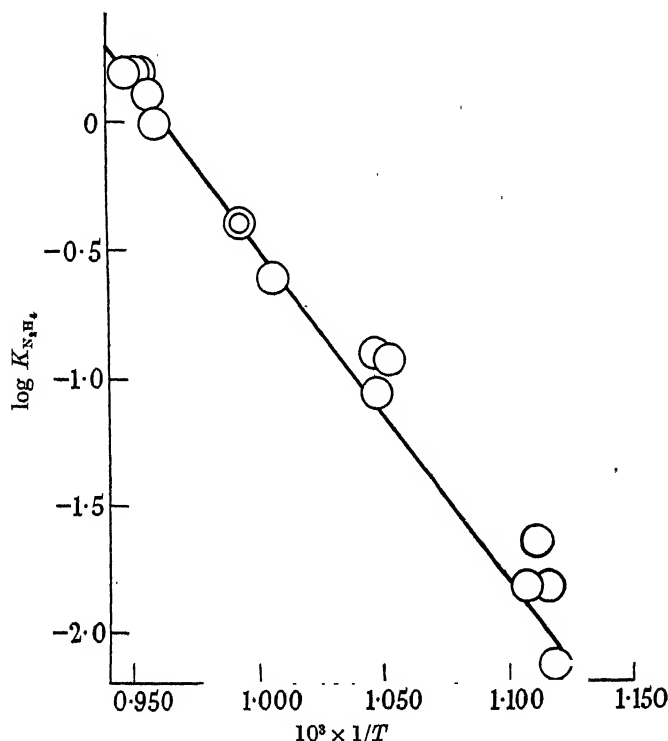


FIGURE 2. $E = 80$ kcal./mole. $\nu = 4 \times 10^{12}$ sec. $^{-1}$. \circ single experiment, \odot average of nineteen experiments.

DISCUSSION

It is now desirable to sum up and review critically the evidence which leads us to the final conclusion that $K_{N_2H_4}$ represents the rate constant of the homogeneous, unimolecular reaction

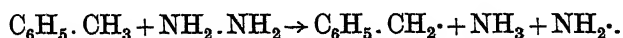


The agreement between the observed and calculated quantities of NH_3 (table 2), and between the observed and calculated quantities of hydrazine introduced (table 4) provides an adequate justification for our method of computation. The latter leads to the conclusion that 1 mole of dibenzyl and 2 moles of NH_3 are formed as a result of the decomposition of 1 mole of hydrazine.

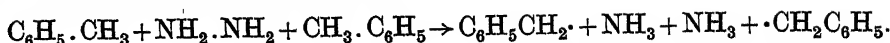
There are three possible schemes which might account for this stoichiometry:

A. The scheme suggested previously consisting of a sequence of reactions (3), (4) and (5).

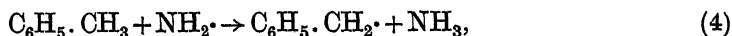
B. The scheme based on a reaction



C. The scheme based on a reaction



The scheme C can be rejected, since the termolecular process is extremely improbable and it would require an increase of $K_{\text{N}_2\text{H}_4}$ proportional to the square of the toluene pressure, in contradiction to our observations (see table 5). The scheme B requires a proportionality between $K_{\text{N}_2\text{H}_4}$ and the pressure of toluene. Table 5 does reveal a slight increase of $K_{\text{N}_2\text{H}_4}$ with increase of the toluene pressure, but this is smaller than would be required by the scheme B. The main opposition to this scheme is provided by the experimentally observed frequency factor and activation energy. To account for the observed values it would be necessary to assume the radius of the activated complex to be at least 700 Å (taking the steric factor as 1). Our present knowledge indicates that the steric factors for these types of reactions are much less than unity (e.g. Glasstone, Laidler & Eyring 1941), which would require an even greater radius for the activated complex. Thus we conclude that the scheme A, i.e.



is the only one which accounts for the observed stoichiometry. Kinetically the process was found to be a homogeneous, first-order gas reaction, with a frequency factor of the magnitude required for a unimolecular decomposition. These characteristics must be those of reaction (3), since reaction (4) should be fast (by analogy with similar processes), and the kinetics of reaction (5) are irrelevant for these considerations.

It could be argued that the experimental activation energy is the sum of the activation energies involved in reactions (3) and (4). This could be true *only* if a very small proportion of the NH_2 radicals react according to equation (4) whereas most of them are destroyed by a wall reaction or by recombination either with themselves or with benzyl radicals. To investigate the last possibility, toluene, recovered from an experiment in which partial pressure of hydrazine was kept very high (0.8 mm. of Hg) and the decomposition was about 85 %, was extracted with water to remove undecomposed hydrazine and was then shaken with water-acetone solution of sodium nitro-prusside (Rimmini test for primary aliphatic amines). The test gave a positive result* (red-violet coloration), and a rough colorimetric estimate determined the quantity of benzylamine as 5 to 10 mole per cent of the available NH_2 radicals. The same test repeated with toluene recovered from an experiment in which the partial pressure of hydrazine was 0.5 mm. of Hg gave only 2 to 5 mole per cent of benzylamine. It is concluded, therefore, that the recombination of NH_2 and benzyl radicals in these experiments is only a minor reaction which can be safely neglected.

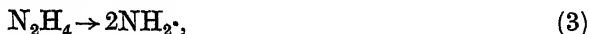
The mutual recombination of NH_2 radicals and their decomposition were studied by investigating the pyrolysis of benzylamine reported in the following paper. It was found that in the presence of toluene there is no mutual recombination of NH_2 radicals, and that their decomposition is only a minor side reaction.

* Rimmini test is given neither by aniline nor by NH_3 or N_2H_4 .

Thus the final conclusion is reached: that under the conditions of the experiments reaction (4) is the only reaction consuming NH_2 radicals which needs to be taken into account for the kinetic calculations.

The dissociation energies of the N-N bond in hydrazine and of the first N-H bond in ammonia

Having obtained the activation energy of reaction



one should be able to calculate the dissociation energy of the N-N bond in hydrazine knowing the activation energy of the reverse process, i.e. the recombination of NH_2 radicals. It is assumed generally that the recombination of radicals or atoms does not involve any activation energy. The case of NH_2 radicals, however, may be an exception. Heitler & Rumer (1931) concluded on the grounds of a theoretical computation that the interaction between two NH_2 radicals leads to a repulsion for large N-N separation. Such a repulsion results in the appearance of a hump in the potential energy curve, corresponding to an activation energy in the process of recombination. However, in the same paper these authors explicitly stated their doubts regarding the physical reality of this hump. They pointed out that their treatment neglected the dipole-dipole attraction, which could well overcome the repulsion effect calculated. It seems, therefore, that it is justifiable to neglect the activation energy of NH_2 recombination, and thus the dissociation energy of the N-N bond in hydrazine can be evaluated as 60 ± 3 kcal./mole. This appears to be a high value, particularly if we recall the 20 kcal./mole advocated by Pauling (1945) as the average bond energy of the N-N bond. It should be noted, however, that Pauling's system of bond energies was formulated on the basis of 170 kcal./mole for the heat of dissociation of N_2 , while the choice of 225 kcal./mole for this heat of dissociation leads to a value of about 43 kcal./mole for the average N-N bond energy (Skinner 1945).

Further support for our value of the N-N dissociation energy in hydrazine can be adduced by considering the dissociation energy of the first N-H bond in ammonia, which can be calculated as follows. The heat of formation of gaseous, anhydrous hydrazine was estimated by Hughes, Corruccini & Gilbert (1939) at +22 kcal./mole, and the heat of formation of NH_3 has been reported as -11 kcal./mole (Bichowsky & Rossini 1936). Using these values, together with $D(\text{NH}_2\text{-NH}_2) = 60$ kcal./mole, a value of +41 kcal./mole was obtained for the heat of formation of the NH_2 radical, from which the dissociation energy of the first N-H bond in ammonia was estimated at 104 ± 2 kcal./mole. This value seems to be reasonable, as it would be expected to lie between the dissociation energies of the first O-H bond in H_2O and the first C-H bond in CH_4 . Dwyer & Oldenberg (1944) estimated $D(\text{HO-H})$ at 118 kcal./mole, whilst Kistiakowsky & van Artsdalen (1944) found the $D(\text{CH}_3\text{-H})$ to be 101 kcal./mole.

TABLE 7

$D(\text{O-H})$	$D(\text{N-H})$	$D(\text{C-H})$
100 kcal./mole	85 kcal./mole	80 kcal./mole
$D(\text{HO-H})$	$D(\text{NH}_2\text{-H})$	$D(\text{CH}_3\text{-H})$
118 kcal./mole	104 kcal./mole	101 kcal./mole

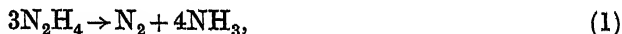
It seems also to be reasonable to expect that the values of the dissociation energies of diatomic radicals OH, NH and CH would follow the same sequence. On the basis of this assumption Glockler (1948) interpolated for $D(\text{N-H})$ a value of about 85 kcal./mole, taking $D(\text{O-H})$ as 100 kcal./mole (Dwyer & Oldenberg 1944) and $D(\text{C-H})$ as 80 kcal./mole (Herzberg 1939). These data are summarized in table 7. Furthermore it is to be expected that the average N-H bond energy in NH_3 would be of the order

$$\frac{1}{2}[D(\text{N-H}) + D(\text{NH}_2\text{-H})] = \frac{1}{2}(85 + 104) \text{ kcal./mole,}$$

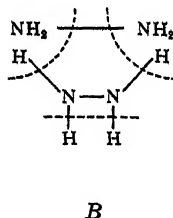
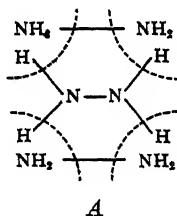
i.e. about 95 kcal./mole. Assuming the heat of dissociation of N_2 to be 225 kcal./mole (Gaydon 1944) one obtains for the average N-H bond energy in NH_3 a value of 93 kcal./mole (Skinner 1945), which is similar to that deduced above. On the other hand, the calculation based on Herzberg's value of the heat of dissociation of N_2 (170 kcal./mole) leads to the average N-H bond energy of ammonia of 84 kcal./mole only. It is concluded, therefore, that the value of $D(\text{NH}_2\text{-H})$, as obtained in this investigation, is in harmony with Gaydon's estimate of the heat of dissociation of N_2 . Very similar arguments, applied by Glockler (1948), also favoured Gaydon's value of the heat of dissociation of N_2 .

Discussion of the heterogeneous decomposition of hydrazine

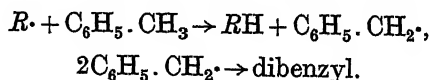
It has already been mentioned that two modes of heterogeneous decomposition of hydrazine have been observed in this investigation:



and



It seems that these reactions occur by the direct decomposition of activated complexes A and B formed on the quartz surface of the reaction vessel. Evidence has been presented in this paper which shows that participation of atoms or radicals in these processes is very unlikely. The formation of radicals or atoms would be expected to produce dibenzyl by a secondary reaction with toluene:



If this were the case the quantity of dibenzyl produced should be increased by packing the reaction vessel, which is contrary to our observations.

Birse & Melville (1940) investigated the heterogeneous decomposition of hydrazine on the walls of the reaction vessel in the presence of para hydrogen. They found that the decomposition did not cause any para-ortho hydrogen conversion. As the con-

version is extremely sensitive to the presence of radicals, these results, in agreement with our observations, point also to the absence of radicals or atoms in the heterogeneous decomposition of hydrazine.*

The formation of the proposed activated complexes, particularly of type A, must correspond to a very low probability factor and, therefore, the decompositions *via* such transition states would only be observed if their activation energies were very small. There was described earlier in this paper the calculation of the activation energy of the overall process at 8 kcal./mole, using the data provided by Birse & Melville. The data presented here have made it possible to estimate the rate constants of the individual reactions (1) and (2). In order to do this one calculates the quantities of N_2H_4 decomposed in each experiment according to equations (1) and (2). The former is given by $3(M_{N_2} - M_{H_2})$ and the latter by $2M_{H_2}$; where M_{N_2} and M_{H_2} denote the number of millimoles of nitrogen and hydrogen collected in each case. The method of calculation is illustrated by means of the examples given on p. 272. The rate constants of the reactions (1) and (2) are given:

$$k_1 = k_{\text{total}} \frac{\text{quantity of } N_2H_4 \text{ decomposed according to (1)}}{\text{total quantity of } N_2H_4 \text{ decomposed}},$$

$$k_2 = k_{\text{total}} \frac{\text{quantity of } N_2H_4 \text{ decomposed according to (2)}}{\text{total quantity of } N_2H_4 \text{ decomposed}}.$$

Table 8 shows the values of k_1 and k_2 calculated in this way for various temperatures. The last column of this table contains the ratio k_1/k_2 . The reproducibility of the results is not very satisfactory but this is not surprising since the walls of the reaction vessel could be expected to change their total activities, and still more their relative activities towards the competing reactions, with time and use. We may recall in this connexion the observations of Gedye & Rideal who found a change in the character of the decomposition of hydrazine on a glass surface which had been sparked previously with a Tesla coil.

TABLE 8

run	<i>T</i> (°K)	<i>K</i> _{total} (sec. ⁻¹)	<i>k</i> ₁ (sec. ⁻¹)	<i>k</i> ₂ (sec. ⁻¹)	<i>k</i> ₁ / <i>k</i> ₂
56	894	0.82	0.72	0.09	8.0
55	896	0.61	0.51	0.08	6.4
15	900	0.61	0.43	0.16	2.7
17-18	903	0.61	0.47	0.12	4.0
20	951	1.10	0.62	0.36	1.7
19	955	0.89	0.47	0.28	1.7
22	955	0.74	0.41	0.25	1.6
27	995	1.6	0.47	0.81	0.6
average at 1008		2.1	~0.8	~0.8	~1.0
5-6	1044	3.5	1.0	1.5	0.7
7-8	1049	3.8	0.6	2.0	0.3
59	1050	4.8	1.7	1.6	1.0
57	1053	4.7	1.1	1.9	0.6
3-4	1057	4.8	0.8	2.4	0.3

* In a private communication Professor H. W. Melville was kind enough to explain to the present writer that the conversion of *p*. H_2 observed at 200° C seemed to be due entirely to the production of the normal H_2 from the hydrazine itself.

The lack of reproducibility makes the estimates of the various activation energies very crude. However, the inspection of table 8 shows quite clearly that reaction (1) corresponds to an activation energy smaller than that of reaction (2), and from the change of the ratio k_1/k_2 with temperature a rough estimate gives $E_2 - E_1 \sim 10$ kcal. The change of k_1 and k_2 with temperature indicates activation energies of the order of ~ 10 and ~ 20 kcal. respectively for processes (1) and (2).

REFERENCES

- Askey, P. J. 1930 *J. Amer. Chem. Soc.* **52**, 970.
Bichowsky, F. R. & Rossini, F. D. 1936 *The thermochemistry of chemical substances*. New York: Reinhold.
Birse, E. A. B. & Melville, H. W. 1940 *Proc. Roy. Soc. A*, **175**, 164.
Bray, W. C. & Cuy, E. J. 1924 *J. Amer. Chem. Soc.* **46**, 858.
Dwyer, R. J. & Oldenberg, O. 1944 *J. Chem. Phys.* **12**, 351.
Elgin, J. C. & Taylor, H. S. 1929 *J. Amer. Chem. Soc.* **51**, 2059.
Gaydon, A. G. 1944 *Nature*, **153**, 407.
Gedye, G. R. & Allibone, T. E. 1931 *Proc. Roy. Soc. A*, **130**, 346.
Gedye, G. R. & Rideal, E. K. 1932 *J. Chem. Soc.* pp. 1160.
Glasstone, S., Laidler, K. & Eyring, J. 1941 *The theory of rate processes*. New York: McGraw-Hill.
Glockler, G. 1948 *J. Chem. Phys.* **16**, 602.
Heitler, W. & Rumer, G. 1931 *Z. Phys.* **68**, 12.
Herzberg, G. 1937 *Chem. Rev.* **20**, 145.
Herzberg, G. 1939 *Molecular spectra and molecular structure*. New York: Prentice Hall Inc.
Hughes, A. M., Corruccini, R. J. & Gilbert, E. C. 1939 *J. Amer. Chem. Soc.* **61**, 2639.
Kistiakowsky, G. B. & van Artsdalen, E. B. 1944 *J. Chem. Phys.* **12**, 469.
Pauling, L. 1945 *The nature of the chemical bond*, p. 53. Ithaca, New York: Cornell University Press.
Skinner, H. A. 1945 *Trans. Faraday Soc.* **41**, 645.
Szwarc, M. 1947 *Nature*, **160**, 403.
Szwarc, M. 1948 *J. Chem. Phys.* **16**, 128.
Szwarc, M. 1949 *J. Chem. Phys.* **17**, 431.
Wenner, R. R. & Beckman, A. O. 1932 *J. Amer. Chem. Soc.* **54**, 2787.

The dissociation energy of the C-N bond in benzylamine

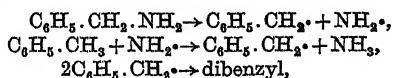
By M. SZWARC, *Chemistry Department, University of Manchester*

(Communicated by M. G. Evans, F.R.S.—Received 19 January 1949)

The kinetics of the thermal decomposition of benzylamine were studied by a flow method using toluene as a carrier gas. The decomposition produced NH_3 and dibenzyl in a molar ratio of 1:1, and small quantities of permanent gases consisting mainly of H_2 . Over a temperature range of 150° (650 to 800°C) the process was found to be a homogeneous gas reaction, following first-order kinetics, the rate constant being expressed by

$$k = 6 \times 10^{12} \exp (59,000/RT) \text{ sec.}^{-1}.$$

It was concluded, therefore, that the mechanism of the decomposition could be represented by the following equations:



and the experimentally determined activation energy of 59 ± 4 kcal./mole is equal to the dissociation energy of the C-N bond in benzylamine.

Using the available thermochemical data we calculated on this basis the heat of formation of the NH_2 radical as 35.5 kcal./mole, in a fair agreement with the result obtained by the study of the pyrolysis of hydrazine.

A review of the reactions of the NH_2 radicals is given.

INTRODUCTION

In view of the complex character of the thermal decomposition of hydrazine described in the preceding paper, it was desirable to carry out some analogous investigations which might be expected to elucidate still further the previously discussed topics. The study of the pyrolysis of benzylamine provided such an opportunity. It made it possible to investigate further the fate of the NH_2 radicals and their reactions with toluene. Moreover, this study furnishes the data for a calculation on a new basis of the heat of formation of the NH_2 radical and the dissociation energy of the first N-H bond in ammonia. The good agreement with the values reported in the previous paper leads to an increased measure of confidence in these results.

EXPERIMENTAL

The apparatus used in the present work was identical with that described in the previous communication. The only difference in the technique, as compared with that employed in the case of hydrazine, was in the method of estimation of the quantities of benzylamine introduced into the system. This was done by direct weighing of the benzylamine reservoir before and after individual experiments. It was noted that benzylamine exposed to air behaved abnormally during the subsequent pyrolysis. To avoid this difficulty the commercial benzylamine was distilled *in vacuo* before use, the middle portion being collected directly into the reservoir. The latter consisted of a small glass bulb provided with a solid key tap and a ground joint. After collecting the required fraction the tap on the receiver was closed

before the pumping was interrupted, thus keeping the distilled benzylamine free from air. During an experiment the reservoir was attached to the apparatus by means of the ground joint and the solid key tap was opened only after the evacuation of the connexions. At the end of each experiment the bulb was immersed in an acetone-solid CO_2 bath, and the solid key tap was closed after the residual benzylamine vapour had been condensed back into the reservoir. Thus it was possible to weigh the reservoir without contaminating its contents with air.

Ammonia and dibenzyl, in molar ratio of 1:1, were the main products of the thermal decomposition of benzylamine in the presence of excess of toluene vapour. Other products of the decomposition consisted of small quantities of permanent gases, their amounts being about 20 % of the amount of ammonia produced. It was shown that the permanent gas contained mainly H_2 (about 70 %).

The toluene recovered from the pyrolysis was tested occasionally for the presence of hydrazine. This was done by shaking it with water, and running the aqueous layer into a solution of HIO_3 which contained some starch. In the presence of hydrazine, HIO_3 is reduced to I_2 and the latter would manifest itself by the appearance of the blue colour, with starch. All the tests proved to be negative showing the absence of hydrazine in the decomposition products.

TABLE 1

	run	T ($^{\circ}\text{K}$)	P_{toluene} (mm. Hg)	$P_{\text{benzylamine}}$ (mm. Hg)	time of contact (sec.)	% NH_3	$k(\text{sec.}^{-1})$	% per- manent gas	% di- benzyl
	52	922	7.1	0.25	0.29	1.7	0.07	28	—
	54	930	6.9	0.15	0.28	2.6	0.09	19	—
	53	937	7.1	0.18	0.29	3.0	0.11	22	—
	49	966	7.2	0.20	0.27	7.0	0.27	22	115
	47	967	15.5	0.12	1.10	31.6	0.34	30	97
	46	968	15.4	0.15	1.10	30.6	0.33	22	110
	48	969	7.3	0.20	0.28	7.6	0.28	22	98
packed	31	1008	7.3	0.18	0.28	30.0	1.25	20	80
	44	1005	15.4	0.08	1.05	72.5	1.2	23	85
	38	1009	6.7	0.15	0.28	27.0	1.1	17	87
	39	1009	6.7	0.16	0.28	25.0	1.0	17	81
	40	1010	6.7	0.16	0.28	27.0	1.1	18	—
packed	29	1010	7.5	0.14	0.28	25.0	1.1	22	93
packed	30	1010	7.6	0.17	0.285	30.0	1.2	20	97
	41	1012	7.3	0.05	0.28	22.5	0.9	31	—
	42	1012	7.5	0.05	0.28	23.0	0.9	25	—
	43	1012	15.4	0.10	1.07	71.5	1.2	29	—
	51	1035	7.4	0.15	0.27	47.5	2.4	20	85
	50	1052	7.5	0.15	0.27	62.5	3.6	20	87
packed	33	1064	7.6	0.17	0.26	67.0	4.3	31	91
	35	1066	7.8	0.17	0.27	87.5	7.7	22	88
	36	1066	7.9	0.15	0.28	86.0	7.0	22	89
packed	32	1070	7.3	0.16	0.255	77.5	5.85	29	78

The results are summarized in table 1; the last column of which gives the quantities of dibenzyl actually isolated, expressed as a percentage of the yields of dibenzyl calculated from the measured amounts of ammonia, adopting the ammonia:dibenzyl ratio as 1:1. Assuming that each mole of ammonia formed corresponds to

1 mole of benzylamine decomposed one finds that in the kinetics of the decomposition of this compound a first-order law obtains. This was established by a fourfold variation of both the partial pressure of benzylamine and the time of contact (see table 1 and figure 1). Moreover, packing of the reaction vessel, which increased the surface by a factor of about $2\frac{1}{2}$, proved that the investigated process was a homogeneous gas reaction (see figure 1). The plot of $\log k$ against $1/T$ over a temperature range of about 140° (from 925 up to 1070°K) is given in figure 1, and from this we estimated the activation energy of the decomposition as $59 \pm 4 \text{ kcal./mole}$ and the frequency factor as $6 \times 10^{12} \text{ sec.}^{-1}$.

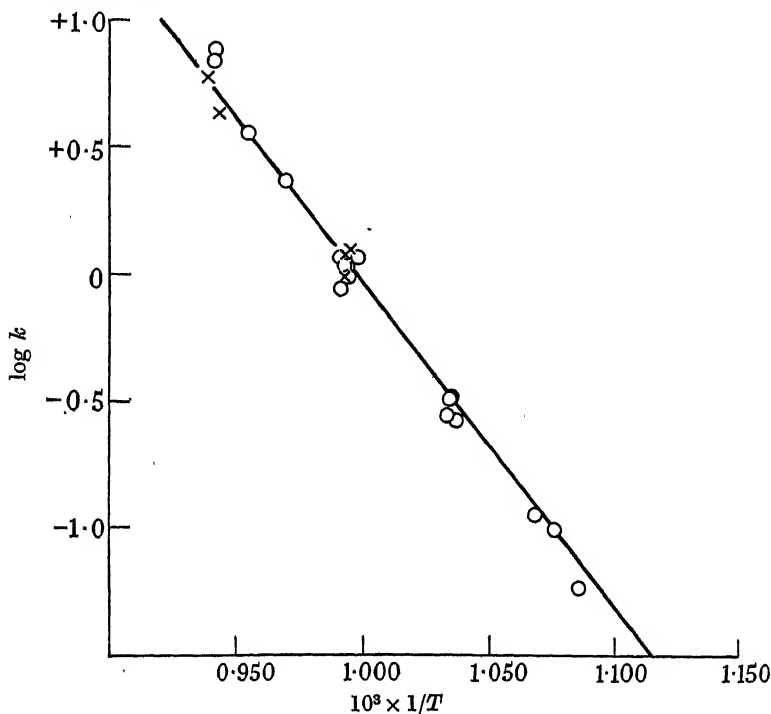
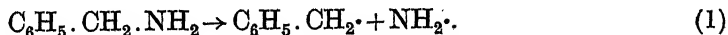


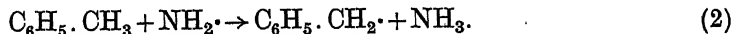
FIGURE 1. \circ unpacked reaction vessel; \times packed reaction vessel.
 $E = 59 \pm 4 \text{ kcal./mole}$. $\nu = 6 \times 10^{12} \text{ sec.}^{-1}$

DISCUSSION

It is beyond doubt that the weakest bond in benzylamine is the C-N bond and, therefore, the initial decomposition step should be dissociation into benzyl and NH_2 radicals,



In an excess of toluene the NH_2 radicals would be rapidly removed from the system by the reaction



The benzyl radicals thus formed would dimerize giving dibenzyl. This proposed mechanism demands the formation of 1 mole of dibenzyl for each mole of ammonia produced, and the results tabulated in table 1 are in agreement with this conclusion.

Reaction (1) is the rate-determining step of the overall process, hence, the kinetics of the formation of ammonia should reveal the characteristics of a unimolecular, homogeneous gas reaction. This conclusion is borne out by the experimental results. The formation of ammonia was found to be a homogeneous, first-order process; moreover, the experimental frequency factor is of the order of the theoretically predicted frequency factor for unimolecular reactions (Polanyi & Wigner 1928). We conclude, therefore, that the observed activation energy of 59 ± 4 kcal./mole is that of reaction (1).

The dissociation energies

Making the usual assumption of zero activation energy for the recombination of radicals one can conclude that the observed activation energy of 59 ± 4 kcal./mole represents the dissociation energy of the C-N bond in benzylamine. Two observations, one reported by Hurd & Carnham (1930) and the other communicated privately by Dr E. Warhurst & Mr B. Gowenlock, indicate that the dissociation energy of the bond in question must be quite considerable.

According to Hurd & Carnham the heating of benzylamine for several hours in a sealed tube at 300°C caused no decomposition whatever, and moreover this compound remained almost unchanged when heated for 54 sec. at 535°C . It may be deduced from these results that the dissociation energy of the C-N bond in benzylamine should be greater than 53 kcal./mole.

Some preliminary unpublished results of Warhurst & Gowenlock suggest a still higher value for the dissociation energy of this bond. These workers pyrolyzed benzylamine using a technique similar to that described by Butler & Polanyi (1943). The extent of the decomposition was measured by the amount of ammonia formed, the latter being estimated by Nessler's reagent. The experiments carried out with N_2 as a carrier gas at 490 and 530°C , using a time of contact of about 0.5 sec., resulted in only slight decomposition of the order 0.1 %. Thus it was concluded by the above authors that the dissociation energy of the C-N bond in benzylamine must be greater than 57 kcal./mole.

Knowing the dissociation energy of the bond in question we are able to calculate the heat of formation of the NH_2 radical using in addition the heat of formation of benzylamine in gaseous state and the heat of formation of the benzyl radical. The latter value is known to be 37.5 kcal./mole from the estimated dissociation energy of the C-H bond in toluene (Szwarc 1948) in conjunction with the relevant thermochemical data, namely the heat of formation of toluene and the dissociation energy of H_2 . Unfortunately, there is no recently measured value for the heat of combustion of benzylamine. The best value seems to be that of Petit (1889) which was chosen by Kharash in his compilation (1929). The heat of vaporization of benzylamine is not available from the literature but by analogy with other amines a value of 11 kcal./mole has been taken for this quantity. Thus one derives the heat of formation of gaseous benzylamine as 14 kcal./mole.

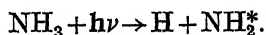
We are now in a position to calculate the heat of formation of the NH_2 radical, using the above data, and the value of 35.5 kcal./mole thus obtained is in good agreement with that derived in the foregoing paper (41 kcal./mole). It should be

noted that the first value involves an error of about 5 kcal./mole, neglecting the uncertainty involved in the heat of formation of benzylamine, whilst the second is uncertain within 2 kcal./mole.

Similar calculations enable us to find the dissociation energy of the C-N bond in methylamine. The heat of formation of the methyl radical is 31 kcal./mole, calculated on the basis of 101 kcal./mole as the dissociation energy of the first C-H bond in methane. Taking the heat of formation of the NH_2 radical as 41 kcal./mole we calculate the dissociation energy of the C-N bond in methylamine as 77 kcal./mole (using the heat of combustion of gaseous methylamine as estimated by Thomsen 1905), or as 81 kcal./mole (using the heat of combustion determined by Muller 1910). It is most unfortunate, that there are no recent data for the heats of formation of various amines.

The fair agreement between the two values of $\Delta H_f(\text{NH}_2)$ as reported in the preceding and the present paper provides additional confirmation for the suggested value of $D(\text{NH}_2\text{-H}) = 104 \pm 2$ kcal./mole. It is desirable, therefore, to review the evidence contributed by other workers to the problem of the dissociation energy of the first N-H bond in ammonia.

The pre-dissociation spectrum of NH_3 observed in the region of 2300 Å by Bonhoeffer & Farkas (1928) indicates that the $D(\text{NH}_2\text{-H})$ must be less than 124 kcal./mole (Bonhoeffer & Harteck 1933). The results of the Hg photo-sensitized decomposition of NH_3 and ND_3 obtained by Melville (1935) lead this author to the conclusion that mercury 1P_1 atoms are quenched by NH_3 to 3P_1 state, and the collision between the latter atoms and NH_3 causes it to dissociate into an H atom and an NH_2 radical. In this case the upper limit for $D(\text{NH}_2\text{-H})$ is brought down to 112 kcal./mole, ignoring the possible formation of HgH . Terenin & Neujmin (1935) investigated the photo-decomposition of NH_3 caused by the Schumann ultra-violet radiation. The decomposition was accompanied by the appearance of the emission band ascribed by these authors to the excited NH_2 radicals. The overall process was assumed to be



The observed energy threshold was 172 kcal. (1650 Å), and the observed excitation energy was 55 kcal./mole. Thus it was concluded that $D(\text{NH}_2\text{-H})$ is less than 117 kcal./mole.

Further information about the value of $D(\text{NH}_2\text{-H})$ could be furnished by the estimation of the activation energies of E_a and E_b of the respective processes



The difference $E_a - E_b = D(\text{NH}_2\text{-H}) - D(\text{H-H})$. E_a was estimated by Dixon (1932) as about 8.5 kcal./mole. This result is based on the collision yield of the reaction between H atoms and ammonia. It seems that roughly the same value should be ascribed to E_b , as the reaction between NH_2 radicals and hydrogen does not take place at room temperature (e.g. Birse & Melville 1940). It is possible, therefore, that $D(\text{NH}_2\text{-H}) \sim D(\text{H-H}) = 104$ kcal./mole.

The fate of the NH₂ radicals

The absence of hydrazine in the products of the pyrolysis of benzylamine indicates clearly that under the conditions imposed in the above experiments the mutual recombination of the NH₂ radicals does not take place. However, the formation of hydrazine by the dimerization of the NH₂ radicals was observed by several workers, and we shall shortly review their results.

Bredig, Koenig & Wagner (1929) reported that hydrazine was formed as a result of an electric discharge through ammonia. The yield of hydrazine increased with the velocity of the NH₃ stream flowing through the discharge tube. Later experiments of Koenig & Brings (1931) revealed that the yield could be increased up to 94 % by cooling the flowing system to -80° C. This dependence of the yield of hydrazine formation on the rate of flow and the temperature seems to indicate that the NH₂ radicals were removed from the discharge tube and subsequently dimerized, most probably on the cooled walls.

The formation of hydrazine in the photo-decomposition of ammonia in a flow system was observed by Gedye & Rideal (1932), who noticed also a very marked increase in the yield of hydrazine when lowering the temperature.

The formation of hydrazine in the photo-decomposition of ammonia in a static system was found beyond doubt only by Weldge & Beckman (1936). These workers noticed that the permanent gases produced by the photo-decomposition of NH₃ contained more than 75 % of H₂ provided the percentage of decomposition was kept very low. They interpreted this as a result of the formation of N₂H₄, and proved the correctness of their assumption by detecting minute quantities of N₂H₄ adsorbed on the walls of the reaction vessel. However, the accumulation of N₂H₄ was prevented by its decomposition in the course of the reaction, and, therefore, the percentage of H₂ in the products was decreasing to 75 % as the percentage of the total decomposition was increasing. This decomposition of hydrazine was the generally accepted explanation for its absence in the products of the photo-decomposition of NH₃ in a static system.

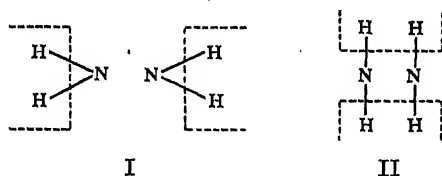
There is a second way of interaction between the NH₂ radicals, namely



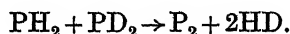
It was stated by Gedye & Rideal (1932) that the role this reaction plays in the process of consumption of the NH₂ radicals increases rapidly with rise in temperature.

Wiig & Kistiakowsky (1932) deduced that reaction (3) cannot occur as a homogeneous gas reaction. Nevertheless, it seems very plausible that this reaction can occur as a heterogeneous reaction taking place on the walls of the reaction vessel. The adsorption of the NH₂ radicals on the surface of silica was demonstrated by Wiig (1935) who found that the quantum yield of the photo-decomposition of ammonia decreased rapidly for very small pressures of NH₃. This effect is due to the diffusion of the NH₂ radicals to the walls, their subsequent adsorption on the silica surface, and finally their recombination with the H atoms. These observations were later confirmed by the investigation of the influence of the size of the reaction vessel on the quantum yield (Wiig 1937).

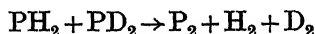
Birse & Melville (1940) developed an elegant technique (rotating sector) which allowed them to study the behaviour of the NH_2 radicals. They proved that the NH_2 radicals have a very long lifetime, being able to survive on the walls of the reaction vessel for a period as long as 70 sec., and found that the decay proceeds according to equation (3). It should be emphasized that this reaction is quite different from the heterogeneous decomposition of hydrazine which leads to the formation of NH_3 , N_2 and smaller quantities of H_2 . Thus the 'normal' N_2H_4 molecule is not an intermediate in reaction (3). It may be, however, that a 'hot' molecule of hydrazine is formed prior to its decomposition according to equation (3). At sufficient low temperatures the rate of the transfer of energy from this 'hot' molecule to the wall is great enough to stabilize it before the decomposition takes place. Thus one can explain the formation of increasing quantities of hydrazine when the temperature is lowered.



There are two mechanisms by which reaction (3) can proceed. One is based on the transition state I, whilst the other assumes the transition state II. The latter is analogous to the transition state *B*, postulated for the heterogeneous decomposition of hydrazine (see the preceding paper). There is no direct information that allows discrimination between these two possibilities. I am, however, inclined to favour the transition state II by comparison with the results obtained by Melville, Bolland & Roxbrough (1937) in their investigation of the PH_2 radicals. These workers had shown that the decomposition of the equimolecular mixture of PH_3 and PD_3 produced H_2 , HD and D_2 . Further, they proved that the mechanism of formation of HD did not involve H or D atoms. Thus the reaction must include the process



Although these results proved the occurrence of the transition state II, the above authors did not exclude the possibility of the reaction



proceeding via transition state I. However, if the latter step is to be included in the mechanism of the reaction, then the quantity of HD produced should be smaller than that expected for the equilibrium mixture of $\text{H}_2 - \text{HD} - \text{D}_2$; although the above authors explicitly stated that the $\text{H}_2 - \text{HD} - \text{D}_2$ equilibrium was completely established.

In the light of all these observations it is not surprising that hydrazine was not detected under the experimental conditions of the present investigation. On the other hand, it could be argued that the small quantities of the permanent gases, formed during the pyrolysis of benzylamine, resulted from the decomposition of the NH_2 radicals as previously discussed. However, this assumption would require

that an increase in the quantity of the permanent gases formed should result from the packing of the reaction vessel which contradicts the actual observations. It was ascertained by special experiments that ammonia is not decomposed in the temperature range used in this study. These results excluded the possibility of producing the permanent gases by the decomposition of NH_3 . There is no satisfactory explanation that can be offered for the process of formation of the permanent gases. In any case it represents only a minor side-reaction, and it does not change the calculated activation energy, as the percentage of these gases in the products remained practically constant for all temperatures. In consequence I have not attempted to investigate further this phenomenon.

REFERENCES

- Birse, E. A. B. & Melville, H. W. 1940 *Proc. Roy. Soc. A*, **175**, 164.
Bonhoeffer, F. K. & Farkas, L. 1928 *Z. phys. Chem. A*, **134**, 337.
Bonhoeffer, F. K. & Harteck, P. 1933 *Grundlage d. photochemie*, p. 128. Dresden: Theodor Heinkopff.
Bredig, G., Koenig, A. & Wagner, O. H. 1929 *Z. phys. Chem. A*, **139**, 211.
Butler, E. T. & Polanyi, M. 1943 *Trans. Faraday Soc.* **39**, 27.
Dixon, J. K. 1932 *J. Amer. Chem. Soc.* **54**, 4262.
Gedye, G. R. & Rideal, E. K. 1932 *J. Chem. Soc.* p. 1160.
Hurd, C. D. & Carnham, F. L. 1930 *J. Amer. Chem. Soc.* **52**, 4151.
Kharash, M. S. 1929 *Bur. Stand. J. Res., Wash.*, **2**, 359.
Koenig, A. & Brings, T. 1931 *Z. Phys. Chem. Bodenstein Festband*, p. 541.
Melville, H. W. 1935 *Proc. Roy. Soc. A*, **152**, 325.
Melville, H. W., Bolland, J. L. & Roxbrough, H. L. 1937 *Proc. Roy. Soc. A*, **160**, 406.
Muller, T. A. 1910 *Ann. Chim. (Phys.)*, (8), **20**, 116.
Petit, P. 1889 *Ann. Chim. (Phys.)*, (6), **18**, 149.
Polanyi, M. & Wigner, E. 1928 *Z. phys. Chem.* **139**, 439.
Szwarc, M. 1948 *J. Chem. Phys.* **16**, 128.
Terenin, A. & Neujmin, H. 1935 *J. Chem. Phys.* **3**, 436.
Thomsen, W. 1905 *Z. phys. Chem.*, **52**, 343.
Weldge, H. J. & Beckman, A. O. 1936 *J. Amer. Chem. Soc.* **58**, 2462.
Wiig, E. O. 1935 *J. Amer. Chem. Soc.* **57**, 1559.
Wiig, E. O. 1937 *J. Amer. Chem. Soc.* **59**, 927.
Wiig, E. O. & Kistiakowsky, G. B. 1932 *J. Amer. Chem. Soc.* **54**, 1806.

The work of the National Institute for Medical Research

BY SIR CHARLES HARRINGTON, F.R.S.

Director of the Institute.

Lecture delivered 10 March 1949—Received 21 March 1949.)

[Plates 9 and 10]

In the National Insurance Act of 1911 there was contained a provision which has proved to be of great importance for scientific work in this country. This provision laid down that the sum of one penny per insured person should be provided from public funds for the purposes of research. The total income resulting amounted to £55,000 in the first year of the operation of the Act, and it was for the administration of this sum and for the decision as to the purposes to which it should be put that the Medical Research Committee was appointed in 1913.

The first report of the Committee, which appeared late in 1915 over the signature of the first secretary, Walter Fletcher, and was submitted to the Chairman of the National Health Insurance Joint Committee, contained a declaration of policy which is of such fundamental importance to the conduct of medical research that it seems to me to be worth quoting in full at this time. In defining their objects the Committee made the following statement:

‘The object of the research is the extension of medical knowledge with the view of increasing our powers of preserving health and preventing or combating disease. But otherwise than that this is to be the guiding aim, the actual field of research is not limited and is to be wide enough to include, so far as may from time to time be found desirable, all researches bearing on health or disease, whether or not such researches have any direct or immediate bearing on any particular disease or class of diseases; provided that they are judged to be useful in promoting the attainment of the above object.’

I shall have occasion later to refer to researches which have been carried out at the National Institute for Medical Research, and which, although they have proved in the outcome to be of importance to medicine, were in their inception purely scientific inquiries. It is clear that only within the framework of the almost academic freedom claimed in the statement of policy which I have read could such researches have been undertaken in an Institute for Medical Research.

Having laid down their general policy in these terms the Medical Research Committee proceeded to consider the steps to be taken for its implementation; the methods decided on were four in number, namely, (1) the employment of full-time investigators in centralized laboratories, (2) the appointment of full- and part-time investigators in hospitals and elsewhere for special researches, (3) the provision of assistance grants to workers in academic institutions engaged in independent researches in the general field of interest, and (4) the maintenance of a statistical department. The individual importance attached to the last of these items must be

ascribed to the obvious bearing of statistical work on the problems raised by national insurance; actually this item was merged with the first, which itself constituted a decision to establish a research institute.

Very soon after the decision had been taken to set up a central research institute, which, incidentally, was to be paid for out of the first year's income of the Medical Research Committee, negotiations were started for the purchase of the Mount Vernon Hospital, Hampstead, with its freehold site and grounds, at the price of £35,000, it being considered, in the words of the report, that these premises offered 'exceptional advantages in amenity and convenience' besides being readily convertible into research laboratories; in addition, it was at this stage intended that part of the building should be used as a research hospital of fifteen to twenty beds.

Whilst these negotiations were in progress a proposal was made to the Medical Research Committee by the Governing Body of the Lister Institute, that the latter should, subject to the agreement of the members and to satisfactory provision for the existing staff, be handed to the nation for use as the central research institute of the Committee. The importance which was attached by the Medical Research Committee to the research hospital part of their project is shown by the fact that the first check in the negotiations with the Governing Body of the Lister Institute arose because the latter possessed no hospital; this difficulty was overcome by the generous offer of Lord Iveagh to build a fifty-bed hospital on a site adjoining the Lister Institute. Moreover, the Medical Research Committee were prepared to take into their employment the whole of the existing staff of the Lister Institute. However, when the proposals were finally submitted to the members of the Lister Institute for approval they were rejected and the scheme came to nothing.

So long as the result of the negotiations between the Medical Research Committee and the Governing Body of the Lister Institute remained in doubt it was naturally impossible to proceed with the work of converting the buildings at Hampstead; the result of this delay was that the outbreak of war in August 1914 found these buildings untouched, and in October of that year it was decided to hand them over to the War Office for use as a military hospital. Thus the control of the Hampstead premises passed out of the hands of the Committee for five years, although much work was done on these premises during that period in which the Medical Research Committee had a direct interest and in which members of their staff took a leading part. When the Mount Vernon Hospital was handed to the War Office it was selected for the study of cardiac disorders—the so-called 'soldier's heart'—and the direction of this work was undertaken by Thomas Lewis, who was seconded to Hampstead from University College Hospital, where he was already working as a full-time member of the Committee's scientific staff. The work on cardiac disorders was continued at Hampstead until 1917 when it was transferred to Colchester, and the Mount Vernon Hospital then became a Central Hospital for Flying Officers; a considerable amount of work was done on the special physiological and medical problems raised by the stresses of flying, and in this again the Medical Research Committee took their part.

In preparation for the staffing of the central institute the Medical Research Committee had already, during 1914, made certain key appointments for the three

main divisions of the work which were at that time envisaged. These appointments were those of Sir Almroth Wright and S. R. Douglas in bacteriology, of H. H. Dale, G. Barger and A. J. Ewins in pharmacology and biochemistry, and of Leonard Hill, Benjamin Moore and Martin Flack in applied physiology; the department of statistics, which was also to be housed in the central institute, was put in the charge of John Brownlee. Since the Hampstead building was not available these workers had to be accommodated elsewhere, and during the war the bacteriologists were based on St Mary's Hospital, although much of their time was spent in France, the pharmacologists and biochemists were at the Lister Institute, and the applied physiologists at the London Hospital; the department of statistics occupied a house in Guilford Street, Russell Square.

The work of the scientific staff of the potential institute during the years 1914-19 was inevitably almost entirely concerned with problems raised by the war; this work I can only mention in passing. Sir Almroth Wright and S. R. Douglas, later joined by L. Colebrook and Alexander Fleming, were engaged on a continuous study of wound infections and dysentery. The pharmacologists and biochemists led by H. H. Dale worked on shock, on trinitrotoluene poisoning and, in conjunction with Clifford Dobell, on amoebiasis; Dale, in particular, also gave much of his time to the control of drugs, especially the arsenical chemotherapeutic agents, and this work was of great importance in relation to later developments of biological standardization to which I shall have to refer. The applied physiologists were occupied with studies of environmental hygiene, and later in the war took part in the research on aviation problems in which Martin Flack was closely concerned.

Possession of the Hampstead building was regained by the Medical Research Committee in June 1919, and work was begun on its conversion into laboratories in the anticipation of occupation by November of that year. Actually, although it was possible to instal the statistics department at an early stage, occupation by the laboratory workers was delayed until April 1920. Even then the equipment was incomplete, and it is recorded that only 'minimum decorative completion' was attempted. During the process of conversion of the hospital into laboratories the nurses' home adjoining the main building was acquired by the Committee and let to the head of the department of biochemistry and pharmacology, H. H. Dale.

In 1920 the Medical Research Committee was reconstituted under grant of a Royal Charter as the Medical Research Council, a body responsible to the Committee of the Privy Council for Medical Research, of which the Lord President of the Council is Chairman. In the Report of the Council for that year the establishment at Hampstead appears for the first time under its present name of the National Institute for Medical Research; it is noteworthy that by this time the plan for the incorporation of a research hospital within the Institute on which so much stress had previously been laid had completely disappeared, although no reference to its formal abandonment is to be found in the Reports of the Council. Reference is made in 1920 to the Institute in the following terms:

'The research work which has been in progress there in all the departments during the past six months has already proved, as the Council think, that the building, though constructed for hospital purposes, is admirably suited for the purposes now

in view. This working experience has fully justified the original choice of this building made by the Medical Research Committee in 1914. With very little structural alteration the rooms have provided convenient and effective laboratories. The workers enjoy almost complete immunity from fog during the winter months, while the building and grounds, though within two minutes walk of the Hampstead Tube Station, are free from the noise of traffic and from electric or other disturbances outside.'

Although my present colleagues who have had to work under conditions of some overcrowding and discomfort during the last few years in the Institute may regard this as a somewhat rosy picture, there is no doubt that in fact the Hampstead building did serve its general purpose admirably for a long time and that the advantages of its situation were not overrated. Nevertheless, unforeseen developments of work with special requirements soon made further provision necessary, and in 1921 the Council purchased a site at Mill Hill where farm laboratories were erected for investigations of viruses which could not be accommodated at Hampstead. Moreover, the value of the Hampstead building itself was greatly increased by the addition in 1928 of an animal house, the cost of which was provided from funds bequeathed to the Council by the late Lord Justice Ronan.

The additional accommodation made available by the construction of the Ronan Building, and by the removal of the Department of Statistics to the London School of Hygiene which occurred in 1927, made the total laboratory facilities at Hampstead adequate until the decision of the Government in 1936 to allocate an additional sum of £30,000 a year to the Council for development of research in chemotherapy raised a new demand for expansion. Since it appeared that such expansion could not be satisfactorily carried out on the Hampstead site the Council made the important decision to erect a new and larger building on their property at Mill Hill to the acquisition of which reference has already been made. This new building was to be designed to house all the existing activities at Hampstead and to provide for the necessary expansion of work in chemotherapy and in other fields of growing importance.

Work on the new building was begun in 1937, and by the summer of 1940 the main structure was complete. At this time, owing to the situation created by the war, the work had to be interrupted and the building was lent by the Council to the Admiralty, by whom it was used as a training establishment for the W.R.N.S. until April 1945. Since then, with many delays, the work of completion of the building for the purposes for which it was originally designed has proceeded, and is now nearly finished, so that from a physical point of view the Institute at present stands on the threshold of the greatest development in its history.

The organization of the Institute in its early days was somewhat peculiar. We have seen that in the appointment of scientific staff the Council had selected leaders in three main branches of laboratory work, in addition to statistics, which stands outside the present discussion. These senior members of the staff each had a department, but there was no Director. Each of the heads of departments, and, indeed, all other members of the scientific staff, were regarded as being directly responsible to the Council through the Secretary.

As might be expected, such an arrangement ultimately proved inconvenient in practice, and in 1928 it was revised by the appointment of H. H. Dale as the Director of the whole Institute. Nevertheless, the idea which lay at the back of the initial organization was an important one; this was that rigid division of the Institute into departments was at all costs to be avoided. The importance of maintaining free intercourse between workers in the different laboratories at the Institute was repeatedly emphasized in the earlier Reports of the Council. It has remained a guiding principle in the administration of the Institute to the present day, and I sincerely hope that it will continue to do so in the future.

Turning now to the actual scientific work of the Institute we have first to notice that the staff which finally assembled at Hampstead in 1920 was different in several important respects from that which had originally been appointed. At the end of the 1914-18 war Sir Almroth Wright and Alexander Fleming returned to St Mary's Hospital, and George Barger was appointed to the Chair of Medical Chemistry at Edinburgh, whilst A. J. Ewins had already taken up industrial research work. When the Institute opened therefore the three laboratory departments were headed by S. R. Douglas, H. H. Dale and Leonard Hill respectively; important accessions to the staff at this time or within the next year or two were W. E. Gye, Clifford Dobell and P. P. Laidlaw in pathology, H. W. Dudley and Harold King in biochemistry and pharmacology, and Percival Hartley, who later took charge of the work on biological standards.

For the first year or two at Hampstead senior members of the staff were chiefly occupied in winding up their researches of the war years, but soon several main lines of new work began to emerge, and it is on the discussion of some of these main lines that I must base my account of the scientific work of the Institute. One of the earliest researches to be begun was that on virus diseases; since this has remained a major theme at the Institute until the present time, and since the development of the work has some points of general interest I propose first to give this some attention.

The importance of virus diseases in human medicine coupled with the ignorance prevailing at that time concerning the viruses themselves were the reasons for the decision to begin systematic investigations of the general problem. The first attempts were directed towards the cultivation of viruses *in vitro*; these having failed, attention was turned to a naturally occurring animal virus disease, namely, dog distemper, which promised to afford a direct experimental approach. With the support of funds raised by the *Field* newspaper, a study of dog distemper was therefore undertaken, the leader of the investigation being the late P. P. Laidlaw with collaboration on the veterinary side from G. W. Dunkin who was appointed to the staff for the purposes of this work. Early on in the study a discovery which later proved to be of major importance was made in the scientific proof that the ferret was susceptible to distemper, a fact which had long been a matter of popular belief. This meant that a convenient experimental animal was available in which the disease could be produced at will, and the conditions were thus provided for a close study of the biological behaviour of the infective virus. In the course of a few years' work sufficient knowledge was accumulated to make possible the

production of an effective vaccine against distemper, with the aid of which, as is now well known, the disease in dogs can be prevented.

The problem of dog distemper having been solved, some time was spent in the study of various other viruses which could be attacked experimentally, including the bacteriophages which were then suspected and are now generally recognized to be viruses parasitizing bacteria. Then in 1933 came the outstanding discovery by Laidlaw, Andrewes and Wilson Smith of the influenza virus. This discovery was made by a return to the technique of the dog distemper work, the ferret again proving to be susceptible to the disease and providing the means for experimental study of the virus. Since that time of course various other techniques for the experimental investigation of viruses have been developed, most notably the method of cultivation in developing eggs first discovered by Goodpasture in the United States and later improved by F. M. Burnet working in the Institute, and there is no doubt that these techniques have greatly extended the possible range of investigations. Nevertheless, it can reasonably be claimed that the work on dog distemper followed by that on influenza really opened up the whole field of precise study of virus diseases in animals and man.

One piece of work which at first formed part of the general virus investigation and which deserves special mention is that of W. E. Gye on the virus factor involved in the production of transmissible tumours. Although all the perplexities of the problem have not been resolved, this work, since greatly extended by Gye himself under other auspices and by many workers elsewhere, together with the independent contribution made to it by C. H. Andrewes, has undoubtedly had a profound influence on cancer research, and it is a matter of pride that it was initiated at the Institute.

The development of knowledge of the biological and epidemiological properties of viruses which resulted from the investigations which I have mentioned naturally led to the desire for more information about their physical properties. Of these the one which clearly first required attention was the actual size of the virus particles, and this problem was attacked in two ways, namely, by the development of quantitative methods of ultrafiltration and by the refinement of optical technique. The work on ultrafiltration resolved itself into a study of the mode of preparation and of the behaviour of standardized collodion membranes having average pore sizes within closely defined limits; by the use of such membranes which were developed by W. J. Elford in the course of a number of years of work and which are now in common use, it was possible to form an estimate of the size of the particles of a virus by ascertaining the minimum average pore size of a membrane which would allow the virus to pass. In this way the sizes of the particles of a number of viruses were assessed, and the values found have in many cases been confirmed by the use of other physical methods since developed.

The particles of most viruses are of course so small that they cannot be observed microscopically with the ordinary microscope; in order to obtain the degree of resolution required it is necessary to use light of wave-lengths shorter than those of the visible part of the spectrum. The use of ultra-violet light offered obvious possibilities in the direction required, and much effort was expended by J. E.

Barnard at the Institute in the development of a satisfactory system of ultra-violet microscopy. The effort put into this work was richly rewarded, not only in the results obtained in the observation of viruses for which it was originally carried out, but in the foundation which it laid for the modern cytological work which, in many laboratories throughout the world, is doing so much to unravel the chemical nature and the biochemical behaviour of the cell nucleus.

The advent in recent years of the electron microscope, with a power of resolution of a different order from that of any optical instrument, has provided another powerful tool for the study of minute objects. Full use has been and is now being made at the Institute of the electron microscopical technique for the examination of viruses. With the aid of new methods of isolation and purification of viruses, and improvements in the method of preparation of specimens, the use of the electron microscope seems indeed to be placing the study of the morphology of viruses upon a new plane.

Whilst great advances have been made in many laboratories in the acquisition of general knowledge about the behaviour and properties of viruses since the beginning of the work at the Institute, it cannot be claimed that progress with the practical treatment of the diseases which they cause has been equally rapid. Indeed, a new advance in this part of the field is the outstanding need of the present. Recent partial successes in the United States, for instance, the discovery of antibiotics such as chloromycetin which cures certain rickettsial diseases and aureomycin which may be active against some true virus infections as well, seem to promise that the advance will come from chemotherapy, and this aspect of the problem is one which is naturally attracting a good deal of attention at the Institute as elsewhere.

Apart from the direct scientific and medical significance of the work on viruses about which I have been speaking, the progress of development of the work and the way in which the problem as a whole has been tackled, have, I think, some points of general interest to which I would like to refer. When the investigation was started it was naturally regarded as essentially a medical problem, and the work was therefore initiated by men whose experience lay in the fields of bacteriology and human pathology. As soon as it appeared that the direct attack was unpromising the direction of effort was diverted to an animal disease, the necessary expert help being obtained by the inclusion of a veterinary surgeon in the group responsible for the research. Again, as the general programme progressed, further additions were made for extensions of the work on the physical and optical sides. In each case the new men who joined the group did so as equal partners, and with freedom simultaneously to develop their own ideas and lines of investigation, which they did with good effect.

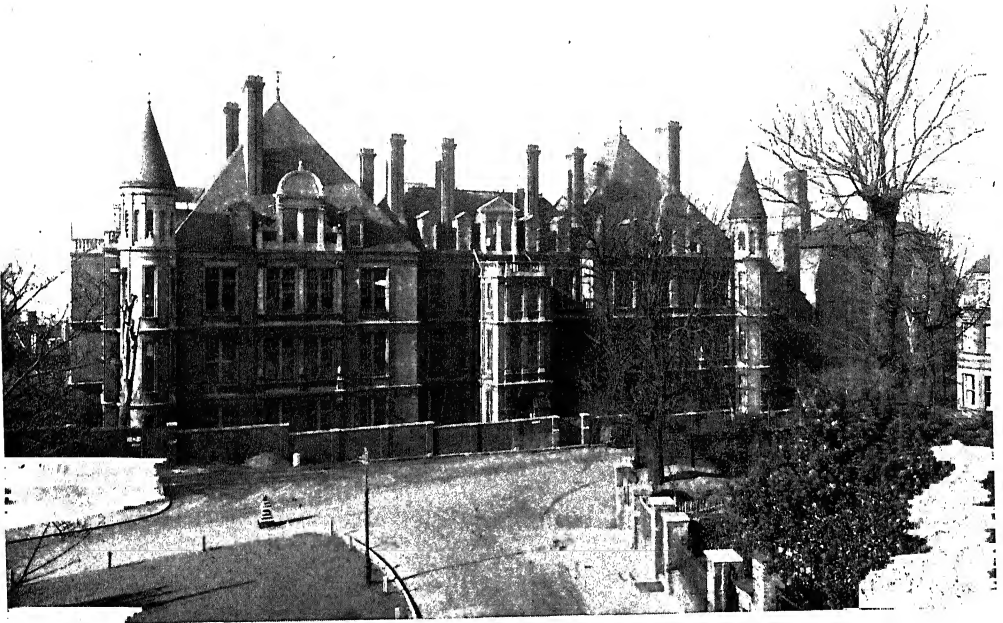
There is nowadays a good deal of discussion as to the part which can usefully be played by teams in research. It seems to me that the history of the virus work at the Institute which I have just described forms a good illustration of effective team work in medical research, although at no time has there been any pretence at the formal constitution of a team. It is perfectly clear that the field of scientific effort covered by the research as a whole has been so wide that no one man unaided could have made much progress. On the other hand, at the outset of the investi-

gation it was quite impossible to predict the devious paths which would have to be followed and the different branches of science into which excursions would have to be made. In these circumstances the course which was actually followed seems to have been the best, the essential feature being readiness at all times to extend the investigation in any desired direction by the addition of the help of an appropriate worker. In this way there is in fact ultimately formed a team of investigators whose efforts are all generally directed to the same end; but the team is never rigidly defined; it grows only in response to the demands of the work, and the individual members retain a considerable degree of scientific independence and scope for initiative. I am myself convinced that, for the organization of work on problems requiring co-operative effort for their solution—and this includes many problems of medical research as it has developed to-day—these are the correct principles.

This, however, is a digression, and I must return to the main theme of the scientific work of the Institute. The next line of investigation to which I wish to refer, particularly because it has been mainly associated with the name of the first Director, to whom it eventually brought the honour of the Nobel Prize for Medicine, is one which in its various ramifications has occupied the attention of a large proportion of the scientific staff of the Institute for many years. In general, it may be defined as the study of the chemical control of bodily functions, and, in fact, it has been mostly concerned with the biochemical and physiological investigation of two compounds, histamine and acetylcholine.

The first of these had long been known as a substance with intense physiological activity, which occurred in nature as a constituent of ergot and as a metabolic product of the action of certain bacteria on the commonly occurring amino-acid histidine. Several of the effects produced by the administration of histamine were closely similar to those observed in anaphylactic shock, and largely as the result of work by Dale it had come to be assumed that the phenomena of anaphylactic shock and of allergy in general were actually due, at least in part, to the liberation of histamine. It also appeared, both from observations by the late Sir Thomas Lewis on the nature of the response of the skin to injury, and from further physiological studies of histamine carried out in the Institute, that liberation of this compound, with its known power of dilating capillaries and causing changes in their permeability, might be responsible for many of the local reactions of tissue to trauma.

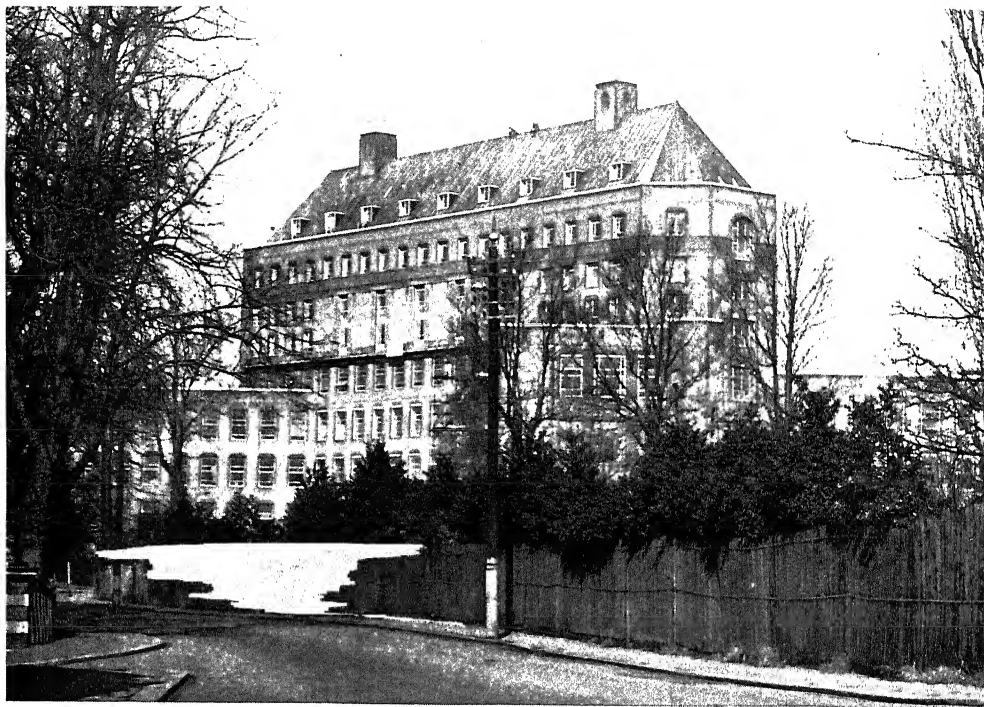
These suppositions naturally implied the existence in normal tissues of a source of histamine, a substance which had so far only been found in the mammalian organism in a situation in which it was presumably a product of bacterial action. In parallel with the physiological studies therefore a systematic search for histamine in various normal animal tissues was begun by H. W. Dudley in the Institute, and led to the discovery that histamine in a bound form was in fact distributed quite widely in animal tissues, in some of which, particularly lung, it occurred in surprisingly large amounts. Thus the objective evidence was supplied which was necessary to establish the supposed role of histamine in producing the phenomena of anaphylactic shock, allergy, and local tissue reaction to injury.



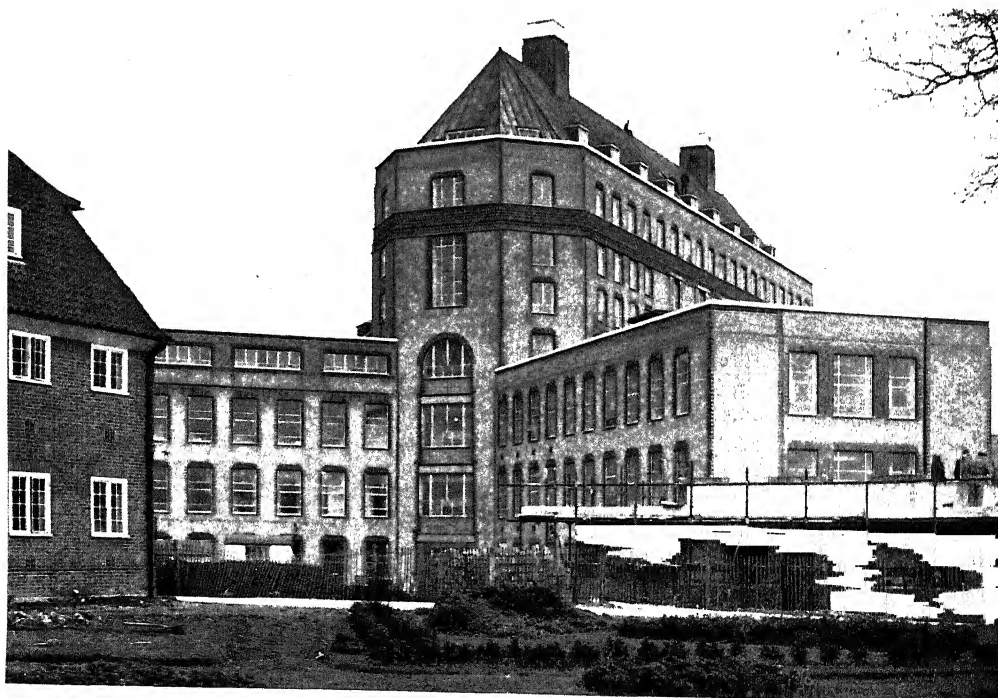
The National Institute for Medical Research, Hampstead. View of front of building.



The National Institute for Medical Research, Hampstead. View of back of building showing Ronan Extension at left-hand side. (Facing p. 300)



New National Institute for Medical Research, Mill Hill. (In course of completion.)
View from the south-east.



Western end of new National Institute for Medical Research, Mill Hill. (In course of completion.)

In the course of the biochemical work on histamine a discovery was made which turned out to be of great physiological importance; this was the observation that extracts of normal spleen contained small amounts of the highly active substance acetylcholine. This compound was hitherto known in nature only as a constituent of ergot. Sufficient information had, however, already been acquired about its physiological actions to suggest that either acetylcholine itself or a substance closely resembling it might be concerned in the production of the effects of stimulation of the parasympathetic nerves. The demonstration of the actual occurrence of acetylcholine as a normal constituent of the body naturally gave a great stimulus to this idea which was actively developed by Dale with a succession of colleagues including J. H. Burn, J. H. Gaddum, W. Feldberg, G. L. Brown and F. C. MacIntosh, in work extending over many years at the Institute and by workers elsewhere. The details of the investigations are not for discussion here; the great generalization to which the work led may, however, be recalled: this is the conclusion that acetylcholine is the substance responsible for the transmission of nervous stimuli from the peripheral ends of parasympathetic nerves to plain muscle, through the ganglionic synapses of the sympathetic nervous system, and at the nerve endings of motor fibres to voluntary muscle; that is to say, it is the chemical transmitter for the effects conveyed by the whole of the efferent fibres of the peripheral nervous system except the postganglionic fibres of the sympathetic system for which a similar part is played by adrenaline. Not only did this generalization form a new chapter in neurophysiology, but the implications of the work as a whole for neurological medicine have been highly significant.

Turning to another series of physiological and biochemical researches, the subject of carbohydrate metabolism in its various aspects is one which has attracted a good deal of attention at the Institute. The interest aroused by the first announcement in 1922 of the discovery of insulin at Toronto caused the Council to send H. H. Dale and H. W. Dudley to Canada in order to report on steps which should be taken to ensure the rapid and proper development of the discovery so far as this country was concerned. The immediately important result of this mission was the acceptance by the Council of the British rights in the patents covering insulin manufacture, and the employment of the control given by these rights to establish British production of insulin on a scientifically sound basis.

A contribution to the method of purifying insulin was made at the Institute which was of some considerable use in the early stages of its manufacture in this country, and the methods for its biological assay were studied in detail in connexion with the control of the quality of the product sold to the public. Apart from this, however, the availability of insulin stimulated interest in the details of the hormonal control of intermediate carbohydrate metabolism. In the course of some years experiments were carried out which revealed several of the underlying facts determining the utilization of carbohydrate in the animal body. A later development arose from the study of the role of hormones produced by the anterior pituitary gland in the control of carbohydrate metabolism, and this eventually led to what is probably the most important discovery in this particular field of work to be made in the Institute, namely, the observation by F. G. Young that it was possible

by purely hormonal means to produce experimentally in dogs and cats a permanent condition of diabetes indistinguishable from the natural disease as it occurs in man.

Before leaving the subject of biochemistry there is one other investigation which I should like to mention, since the results of the work will always be associated with the name of the Institute; this is the research work which led to the identification of one form of vitamin D—that which is now known as vitamin D₂. The whole study was again pre-eminently an example of co-operative research; it started with an inquiry into the nature of the process by which vitamin D activity was conferred on sterols by irradiation with ultra-violet light; this inquiry revealed the identity of the provitamin with ergosterol and opened the way for experiments on the direct irradiation of this compound which ultimately led to the isolation of pure vitamin D₂, or, as it was called, calciferol. This work is a further example of a research carried out by a team which grew during its progress; having been initiated by O. Rosenheim and T. Webster it was continued with collaboration, among others, from R. B. Bourdillon on the physical side, and in its organic chemical and biochemical aspects from R. K. Callow and Miss H. M. Bruce.

Apart from this detailed study of one vitamin, and a contribution on the nutritional significance of phytic acid in cereals, nutritional research has not been undertaken to any considerable extent within the Institute itself; on the other hand, in the associated Nutrition Laboratory a continuous programme of such work has been maintained, amongst the results of which we may recall the investigation of abnormalities of bone growth associated with vitamin A deficiency which formed the subject of Sir Edward Mellanby's Croonian lecture to this Society in 1943, and the recently much discussed discovery of the toxic effects of 'agenized' flour.

The Medical Research Council have always shown themselves to be fully conscious of the contribution which chemistry can make to medical research. We have already seen the part which chemists and biochemists have played as collaborators in several of the major themes of research which I have mentioned. They have, however, also made their own independent contributions, and particularly in the development of chemistry in relation to pharmacology and chemotherapy. We may recall, for instance, two advances which have resulted from a large amount of work on alkaloids. The first of these arose out of clinical observations by Chassar Moir at University College Hospital; it led to the isolation by H. W. Dudley from ergot extract, at a time when it was thought that everything was known about the alkaloidal content of this material, of a new alkaloid which actually turned out to be the really important constituent of ergot from the point of view of its practical use in medicine. The second is the work of Harold King on curare alkaloids which has led to the isolation and chemical identification of the pure alkaloid D-tubocurarine which is individually responsible for the typical physiological action long associated with curare. This work also has had its therapeutic application in the extensive use which has been and is still being made of D-tubocurarine as an adjuvant in anaesthesia to facilitate the task of the surgeon by providing more complete muscular relaxation. As an extension of the investigation of curare an interesting advance

has been made quite recently which again may have practical application in medicine; this consists in the developments by chemists and pharmacologists working in collaboration, of a series of simple bisquaternary salts, one of which promises to be therapeutically equal to or better than D-tubocurarine, whilst others may find therapeutic applications of quite a different kind.

So far as chemotherapy is concerned, this has always been a primary interest in the organic chemical laboratories of the Institute. It cannot be denied that the workers in these laboratories have had their full share of the disappointments which seem to be inevitably associated with this most exacting and tantalizing form of scientific effort. Chemotherapy is a subject in which, at least until recent years, there has been little or no basis of theory upon which a logical attack of a problem can be planned. This being so, the probability of success has necessarily been determined by the quantitative magnitude rather than the qualitative nature of the experimental investigation. In comparison with the effort which has been put into chemotherapy by large industrial organizations, that at the Institute has, of course, been extremely small; nevertheless, it has not been without its successes, and in one case at least, namely, in the part played in the development of the amidine drugs for the treatment of kala azar, the practical outcome has been significant. At any rate there is every intention that chemotherapy in all its aspects shall continue to be a principal feature of research work at the Institute, particularly in relation to the outstanding problems of the chemotherapeutic treatment of tuberculosis and of true virus diseases.

The staff of the chemical laboratory at the Institute have also been responsible for a major theoretical contribution to chemistry. This was the suggestion which was advanced in 1932 by Rosenheim and King for the structure of cholesterol and which brought about a reorientation of the chemistry of the whole group of steroids—a group that includes so many compounds of biological importance.

One of the more remarkable developments in medical and biological research in recent years is the increasingly important part which is played by physical methods. So far as the Institute is concerned the illustration of this, which is very much in my mind, is that the plans which had been made immediately before and during the early part of the recent war for the occupation of the new building have had to be fundamentally revised to provide the new facilities which are demanded by this particular branch of the work. We have already seen how even in the early stages of the virus work special physical and optical techniques had to be employed in order that the progress of the work should be maintained, but at this time the techniques in question were relatively simple, and modest in their demands for space and personal effort. Now it is entirely necessary to provide facilities for much more complicated procedures; such as phase contrast microscopy, electron microscopy, electrophoresis, high-speed centrifugation and ultrasonics. In addition to all this the introduction of the use of isotopes as valuable tools in biological research and particularly the almost unlimited availability of radioactive isotopes for this purpose has necessitated the undertaking of an amount of purely physical work which would have seemed fantastic for an institute of medical research even as little as 10 or 12 years ago.

All the physical techniques which I have mentioned are now available in the Institute, where they are used not only by workers within the Institute itself, but for those engaged in medical research problems in other institutions, who require to use special physical methods for their work but have not themselves access to the necessary apparatus and skilled help. This state of affairs is particularly noticeable in respect of the isotope work; on the one hand, the Institute mass spectrometer is made available for analyses for medical research workers in Medical Research Council Units or University laboratories who are using stable isotopes in their experiments; on the other, the staff of the Biophysics division have the task, at least so long as the present position of dependence on American supplies persists, of receiving, assaying and distributing the radioactive isotopes required for the many research projects which are being carried on throughout the country with the aid of these materials, in the medical and biological field. Naturally the ready availability of this variety of physical techniques within the Institute is itself a stimulus both to the further development of the methods themselves and to their application wherever this seems likely to aid the solution of a problem.

To return to the biological work of the Institute I must first refer to that of the Department of Applied Physiology. It will be remembered that this department, under the leadership of Leonard Hill, was one of the main divisions at the outset of the activities of the Institute, and it remained so until Hill's retirement in 1930. The first main theme of work of this department was the study of environmental factors in their relation to health and activity, and among the more important results of this work was the development of the kata-thermometer, an instrument which has been widely used for many years and is still employed with good effect in the assessment of conditions of ventilation. Much time was also spent in the study of the physiological effects of ultra-violet light, and this work linked up with the inquiry already described which ultimately led to the isolation of calciferol. In still another aspect of its work the department was concerned with the physiological effects of high and low atmospheric pressures, the former particularly in relation to the problems of deep-sea diving. It is of some interest to note that after a lapse of 20 years there is once more to be a division of Applied Physiology within the Institute, the staff of which will again be engaged in the study of the effect of environmental conditions on various human activities.

Shortly after the department of applied physiology ceased to exist as a separate division the biological work of the Institute was extended in another direction by the establishment of a division of endocrinology under A. S. Parkes. This became a widely recognized centre for the study of the sex hormones in particular, and was responsible not only for advances in the knowledge of the fundamental physiology of these compounds, but for practical discoveries which facilitated their therapeutic use. The interest of the workers in this department in the anterior pituitary gland had also much to do with the development of the experiments of F. G. Young on diabetes which have already been mentioned. As time has gone on the activities of this division have extended beyond the limits of endocrinology, and its field of work is now better covered by the more general title of experimental biology, the current interest being chiefly in problems of fertility.

I come now to a part of the work of the Institute which is not confined to pure research, namely, the maintenance of biological standards. I have already referred to the work which had to be done during the 1914-18 war on the control of drugs, particularly arsenicals, and for which H. H. Dale was chiefly responsible. This work was the foundation of an enduring interest in problems of standardization of physiologically active compounds whose potency had to be assayed by biological means. Many substances which have important uses in therapeutics are not, particularly when they are first introduced, sufficiently well defined to be properly assayed by chemical and physical tests. Such a substance therefore has to be tested for potency by a biological test for its specific activity, and all biological tests are subject to an inherent variability which makes them unsuited in themselves as a basis of quantitative measurement. The only way out of the difficulty is to set up a permanent standard preparation of the substance in question and to define biological activity in terms of a known weight of this standard preparation. If new preparations are then tested biologically in direct comparison with the original standard or with a substandard prepared therefrom their activity can be expressed in terms of a known weight of a stable substance, instead of in terms of the behaviour of an animal on a particular day. It is on this fundamental principle that the whole work of biological standardization is based.

In the Report of the Council for 1920 we find the remark that at every point the work of the departments at the National Institute for Medical Research has concern with biological standards, and it is noted that a department of biological standards had been set up at the Institute under H. H. Dale with appointment of an additional member of the staff for special work in connexion with it.

The main types of substances for which biological standards are needed are serological preparations such as antitoxins, vitamins, hormones, and certain drugs such as digitalis. In the early years at the Institute much work was done on the preparation of standards for diphtheria and tetanus antitoxins, digitalis, and posterior pituitary gland. In 1923 steps were taken to correlate the work which had been done up to that time on biological standards in this and other countries, and international agreement was obtained for the establishment of standards for diphtheria antitoxin and posterior pituitary gland. In the same year the League of Nations Health Organization assumed responsibility for biological standardization as a whole; the international status of the work has been maintained from that time; after the recent war responsibility was taken over by an interim commission and it is now passing into the hands of the World Health Organization. In 1922 began the long association with work on standards of Percival Hartley, who was appointed to the staff in that year and later directed the Department of Biological Standards within the Institute until 1946.

A further event which was of importance to medicine in this country occurred in 1925 when the Therapeutic Substances Act was passed. This Act provided for the official control of therapeutic substances whose potency had to be determined by biological means and in the exercise of this control a new responsibility was placed upon the staff of the Department of Biological Standards at Hampstead. In the schedules of the Act the units of activity of each of the therapeutic preparations

are defined in terms of the standards which are preserved in the National Institute for Medical Research. The Department there is responsible for the issue of samples of the appropriate standards to approved manufacturers who may require them, and, furthermore, for ensuring that the products placed on the market by these manufacturers satisfy the requirements laid down in this Act.

As time has gone on and as new biological products have been brought into therapeutic use, the work of the Department of Biological Standards has considerably increased. There are now some thirty-seven standards preserved at Hampstead, of which thirty-five have international status, mostly in the groups of vitamins, hormones and serological preparations, with the addition recently of antibiotics such as penicillin and streptomycin. Samples of these standards are, as I have said, issued to manufacturers, who customarily prepare their own substandards for the control of their products, and with whom the closest touch is maintained on all technical questions arising from the work. The standards are also available to research workers in all countries. In many countries National Control Centres have been established, and where these exist distribution of the ultimate reference standards from Hampstead to the countries concerned is effected through them.

The actual way in which the work on standards is done at the Institute is perhaps worth noting. The amount of technical work required in connexion with the establishment of new standards may be considerable, but the demands for such work are inevitably intermittent, and the experimental technique required may fall in any one of a large variety of fields. For this reason it has not been thought wise to establish a large department devoted exclusively to work on standards. The director of the Department has the responsibility for the co-ordination of the whole of the work, and for expert advice and assistance he is able to call on his colleagues in the other departments of the Institute who possess the relevant special knowledge. This arrangement has the great advantage that the necessary technical work becomes an occasional demand on the time of members of the scientific staff who are otherwise engaged on their own research problems, instead of being a full-time occupation of a somewhat restrictive character. In actual fact many members of the staff who have given most help with standards, including those in the Department itself, have themselves been responsible for some of the more important contributions in research. One need only recall the researches of Hartley himself on general problems of immunology and those of Bruce White on the cholera vibrio to realize the truth of this statement, whilst among workers less closely connected with the Department, the physiologists and the endocrinologists have given much valuable assistance to this work.

Biological standards, besides forming an essential part of the system of control of biological therapeutic products, offer a substantial service to research; moreover, the work which arises in connexion with them itself stimulates research particularly in the important field of biological assay. The existence of the Department is furthermore of great value to the Institute itself; through the department valuable contacts are maintained with the research laboratories of industrial firms, and, since the preparatory work for the establishment of new standards is always done so far as

possible with international co-operation, contacts are also close with research workers in other countries.

There is no part of a research institute which is of more importance to the general progress of the work than the Library; I cannot conclude my description of the National Institute for Medical Research without mentioning that as a result of the liberal policy of the Council and of the devoted efforts of the librarians the Institute now possesses a library which, having started from very small beginnings, has become one of the best medical research libraries in the country. The facilities of this library are of course not restricted to workers within the Institute itself, but are available to other workers under the Council, and, by special arrangement, to medical research workers elsewhere.

This brings me to the end of my outline of the main activities of the National Institute for Medical Research. It has been an account which makes no claim to completeness, but in which selected themes have been briefly discussed as illustrations of the type of work which is done and of the main trends of development. The story which I have told is one of an expanding effort, the expansion relating both to the numbers of the scientific staff and to the range of scientific work which is covered. This expansion has been continuous, not even being completely interrupted by the recent war. During the war the activities of the Institute were fully maintained, although there was inevitably and properly a considerable temporary redirection of the research effort to problems of immediate importance; among these may be mentioned in passing the work of the physiologists on underwater physiology in connexion with submarine warfare, work on measures of protection of troops against flamethrower attack, and studies of protective measures against rickettsial diseases of military importance. With all this it was possible to keep pace with new technical developments and to maintain a background of fundamental research so that resumption of normal activities on an increased scale was not too long delayed.

Looking in more detail at the process of expansion of the work at the Institute from the early days of the three laboratory departments to the present very much larger and more complex establishment we see an increasing emphasis on biochemistry and biophysics on the one hand together with a broadening of the biological research through endocrinology to more general aspects of experimental biology on the other. In these respects the development is a true reflexion of the general trend during the past thirty years of medical research in the laboratory, which must make ever wider demands on the whole field of scientific effort if it is to progress. The very increase in the range of work which has to be covered, with its accompaniment of multiplication of departments with differing immediate interests, adds weight to the principle laid down by the Council in the early days of the Institute that there must be the freest contact between the different laboratories.

If the organization which I have depicted seems to some to be ill-defined and loose in structure, this is because of the pre-eminent importance which is attached to the principle of flexibility and avoidance of departmentalization. It is, indeed, in my view only with the strictest adherence to such a principle, and within the

framework of the early declaration of research policy which I quoted at the beginning of this lecture, that we can hope for the Institute to continue to fulfil its proper function of producing new knowledge in science and medicine, and in doing so to live up to the achievements of the past, of which it may reasonably be proud.

Kinetics of the base-catalyzed halogenation of some ketones and esters

BY R. P. BELL, F.R.S., E. GELLES AND EVA MÖLLER

(Received 1 March 1949)

Kinetic measurements have been made at 25° C on the halogenation of benzoylacetone, acetylacetone, ethyl acetoacetate, ethyl α -bromoacetoacetate and diethyl bromomalonate. A method is described for analyzing the kinetic data and obtaining the rates of substitution of successive halogen atoms without isolating the partly halogenated derivatives.

The reaction velocities measured are all independent of the halogen concentration, and represent the rates of ionization of the ketonic substances in presence of basic catalysts. The results obtained conform in general to the regularities previously found for this type of reaction, but anomalies are found in some instances; these are related to the interaction of large substituent groups in both substrate and anion catalyst. The effect of bromine substitution in increasing the reaction velocity is shown to decrease as the reactivity of the ketone increases, and this is explained in terms of the charge distribution in the anion of the substrate.

INTRODUCTION

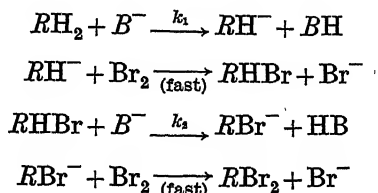
The original object of this work was to extend to benzoylacetone the investigations of base-catalyzed prototropy which have already been made with a series of ketonic substances (Bell & Lidwell 1940; Bell 1943; Bell, Everett & Longuet-Higgins 1946; Bell, Smith & Woodward 1948). However, measurements with this ketone led to two more general points of interest. In the first place a new method was devised for obtaining the rates of substitution of successive halogen atoms without isolating the partly halogenated derivatives, and this method was used to re-examine the bromination of acetylacetone and acetoacetic ester. In this connexion measurements were also made on the bromination of α -bromoacetoacetic ester and bromomalononic ester. In the second place it was found that certain anion bases showed an unexpectedly great catalytic effect in the bromination of benzoylacetone and bromoacetylacetone, and the same anions were therefore investigated as catalysts in the bromination of acetoacetic ester.

MEASUREMENTS WITH ACETYLACETONE AND BENZOYLACETONE

(a) *Treatment of successive reactions*

As with all the substrates mentioned in this paper, the rate of bromination of these two ketones is independent of the halogen concentration, and thus depends on the rate of ionization of the ketone in presence of the basic catalyst, usually an anion

B^- . In each case two of the hydrogens of the group $-\text{CO} \cdot \text{CH}_2 \cdot \text{CO}-$ are eventually substituted by bromine, and the general kinetic scheme may be written as follows:



In this scheme RH_2 and RHB represent the keto-forms of the ketonic substances concerned, and k_1 and k_2 are first-order velocity constants, the catalyst concentration remaining effectively constant. In previous work with acetylacetone (Bell & Lidwell 1940) it was assumed that $k_2 \gg k_1$, leading to simple first-order kinetics, but we shall see that this assumption is not justified. The more general case ($k_1 \sim k_2$) has been treated by Pedersen (1933), who has also taken into account the equilibrium fraction of enol, e , present in the original ketone solution, which will be brominated instantaneously. The resulting kinetic expression is

$$2(r-1)(c - \frac{1}{2}x)/c = (1-e)(2r-1)10^{-k_1 t} - (1-re)10^{-k_2 t}, \quad (1)$$

where c = initial total ketone concentration,

x = equivalents of bromine present at time t in $\text{RHB} + \text{RBr}_2$,

k_1, k_2 = first-order constants in terms of decadic logarithms,

$r = k_2/k_1$.

This expression reduces to a simple first-order change only if $k_2 \gg k_1$, or $k_1 = 2k_2$. In other cases it should be possible in principle to deduce the values of k_1 and k_2 from the observed course of the bromination, but in practice this can only be done by a laborious process of trial and error, and the resulting values of k_2 are not of high accuracy (cf. Pedersen 1933, 1934).

If it is possible to study separately the bromination of RHB , thus obtaining k_2 , then k_1 can be obtained fairly easily from the observed bromination kinetics of RH_2 . This is most readily done by a process of interpolation. A value is assumed for r ($= k_2/k_1$), and values of the expression

$$X = \log_{10} \{ 2(r-1)(c - \frac{1}{2}x)/c + (1-re)10^{-k_2 t} \} \quad (2)$$

are plotted against t . If the correct value has been chosen for r , the resulting plot will be a straight line of slope k_1 ($= k_2/r$), while an incorrect value for k_1 will give a line which may be noticeably curved, and will in any case have a slope differing from the value assumed for k_1 . By taking a series of values for k_1 the correct value can be obtained by interpolation. (If k_2 is appreciably greater than k_1 , an approximate value for k_1 can be obtained by plotting $\log_{10}(c - \frac{1}{2}x)$ against t .)

This procedure is satisfactory only if r differs considerably from unity. If we put $r = 1$ in equation (2) the first term vanishes, and for any value of r close to unity the plot of X against t is insensitive to the experimental values of x . In the range $0.7 < r < 1.3$ the best method is to write $r = 1 + \eta$ in equation (1) and to expand in

powers of η . Correct to terms in η^2 (which is sufficient for present purposes) equation (1) then becomes

$$(c - \frac{1}{2}x)/c = 10^{-k_1 t} \{1 - \frac{1}{2}\epsilon + 1.15k_2 t(1 - \epsilon - \eta + \eta^2) - 1.33k_2^2 t^2 \eta(1 - \epsilon - 2\eta + \epsilon\eta) + 1.02(1 - \epsilon)k_2^3 t^3 \eta^2\}. \quad (3)$$

Defining an expression Y by

$$Y = \log_{10} \{(c - \frac{1}{2}x)/c\} - \log_{10} \{1 - \frac{1}{2}\epsilon + 1.15k_2 t(1 - \epsilon - \eta + \eta^2) - 1.33k_2^2 t^2 \eta(1 - \epsilon - 2\eta + \epsilon\eta) + 1.02(1 - \epsilon)k_2^3 t^3 \eta^2\}, \quad (4)$$

Y is plotted against t for a series of values of η . When the right value of η is chosen, the resulting plot will be a straight line of slope $k_1 = k_2/r = k_2/(1 + \eta)$. Examples of these procedures are given later in the paper.

The monobromo-derivatives of acetylacetone and benzoylacetone are not well characterized, and no attempt was made to prepare them in a pure state. Their bromination in solution can, however, be studied by the following procedure. Both acetylacetone and benzoylacetone are sufficiently strong acids ($\text{pK} \sim 9$) to be converted almost completely to the enolate ion by a slight excess of alkali. Both the ion and the enol are known to react instantaneously with bromine, so that on adding an excess of bromine solution (if necessary containing buffer solution) to a solution of ketone containing a slight excess of alkali the monobromoketone is formed instantaneously and is then brominated further at a measurable rate, enabling k_2 to be measured directly.

(b) *Experimental procedure*

Acetylacetone was redistilled twice and had b.p. 139°C . Benzoylacetone had m.p. 57°C , unchanged by recrystallizing from aqueous alcohol. The acids used for preparing buffer solutions were either A.R. samples or were purified by recrystallization or distillation, their purity being checked by titration against standard sodium hydroxide solution which was in turn standardized against constant-boiling hydrochloric acid. The glycollate buffer solutions were made from recrystallized sodium glycollate. All solutions were made up with boiled-out distilled water using grade A volumetric apparatus. The automatic pipette used for delivering bromine solutions was of the type described by Bell, Lidwell & Vaughan-Jackson (1936).

The kinetic measurements were carried out at $25 \pm 0.01^\circ\text{C}$, and the ketone concentration in the reaction mixtures was about 4×10^{-4} throughout. Owing to the limited solubility of benzoylacetone in water some of the measurements with this ketone were made in solutions containing 3 % of alcohol or acetone, it being found that 7 % of organic solvent had no detectable effect upon the reaction velocity. In measuring the two-stage bromination process, 20 ml. of a solution was prepared containing ketone, buffer, and enough sodium chloride to make the ionic strength of the final solution up to 0.1. Approximately 1 ml. of 0.5M-bromine in potassium bromide solution was transferred by an automatic pipette into a thin-walled bulb which rested inside the flask containing the ketone solution. When the whole had reached thermostat temperature the reaction was started by breaking the bulb

while shaking the flask and simultaneously starting a stop-clock by depressing a key with the foot. After a suitable interval the reaction was stopped by adding a few drops of 10 % allyl alcohol solution, and the clock simultaneously stopped, recording the reaction time to the nearest 0.1 sec. The mono- and dibromoketone were then converted into iodine by adding a crystal of potassium iodide and 2 ml. 2*N*-hydrochloric acid; liberation of iodine was complete within a few minutes, and it was estimated by titration with *N*/100 or *N*/200 thiosulphate using a microburette and sodium starch glycollate indicator (Peat, Bourne & Thrower 1947). A series of experiments (normally ten to fifteen) with different times was carried out for each reaction mixture. The second stage of the reaction was followed in a similar manner, except that the ketone solution contained slightly over 1 equivalent of sodium hydroxide. The anion catalyst was contained in the same solution, and the admixture of the bromine solution (containing, if necessary, some strong acid) produced a buffer solution of suitable ratio. Experiments with varying concentrations of strong acid (and hence no basic catalyst other than water) gave identical results. There is therefore no detectable catalysis by acids, and the buffer ratio in the experiments with buffer solution is not in itself important, though it must be known in order to calculate the correction due to hydrolysis of the bromine (see below).

The above technique gave reliable results with reaction times as low as 2 sec., and the half-times of the reactions studied varied between about 10 and 150 sec. The observed titres for a typical experiment (the water reaction for acetylacetone) are shown in figure 1. It will be seen that the titres for long reaction times do not approach closely to the theoretical end-point, and in fact show a tendency to decrease again. Benzoylacetone behaved in the same manner, though to a less marked extent, and a similar decrease in reactive halogen content has been observed in other cases (Bell & Lidwell 1940; Pedersen 1933, 1934); it is probably due to hydrolytic fission of the halogenated ketone. This complication does not affect the earlier part of the reaction, and in calculating velocity constants only the first half of the reaction was used for acetylacetone, and the first two-thirds for benzoylacetone.

In the overall bromination experiments the extrapolated initial titres agree closely with the equilibrium enol contents of the aqueous solutions (17 % for acetylacetone, 34 % for benzoylacetone; see Eidinoff 1945), assuming one bromine atom to be introduced instantaneously. When the second stage of bromination was being studied the initial value corresponded closely with 100 % enol, and was independent of the interval (up to 20 min.) which elapsed between making the solution alkaline and adding the bromine solution, thus showing that the alkali had not caused any appreciable fission of the ketone. In either case, if r_{∞} is the calculated titre for complete bromination, and r_t the observed titre at time t , it was found that $\log_{10}(r_{\infty} - r_t)$ is a linear function of t . This is illustrated by figure 2, which gives four specimen plots for experiments with acetylacetone in monochloracetate buffers. The arrows indicate calculated initial titres. For the second stage of the reaction the slope of these plots gives k_2 directly, and the values of k_2 thus obtained are a linear function of the anion concentration, as shown in figure 3. For the overall reaction, however, the slope has no simple significance, and the data must be treated by the method described in the last section. The values of the ratio r differ sufficiently from unity to justify the

application of the first method described (cf. equation (2)). As an example of the results obtained the data for monochloracetate solutions are given in table 1. In this and in all subsequent tables the values of c , the anion concentration, have been

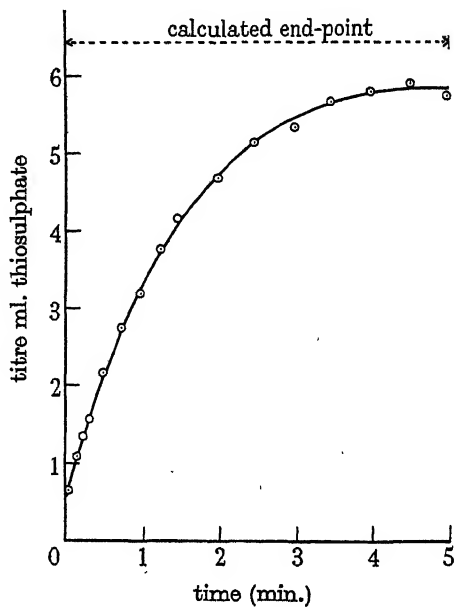


FIGURE 1

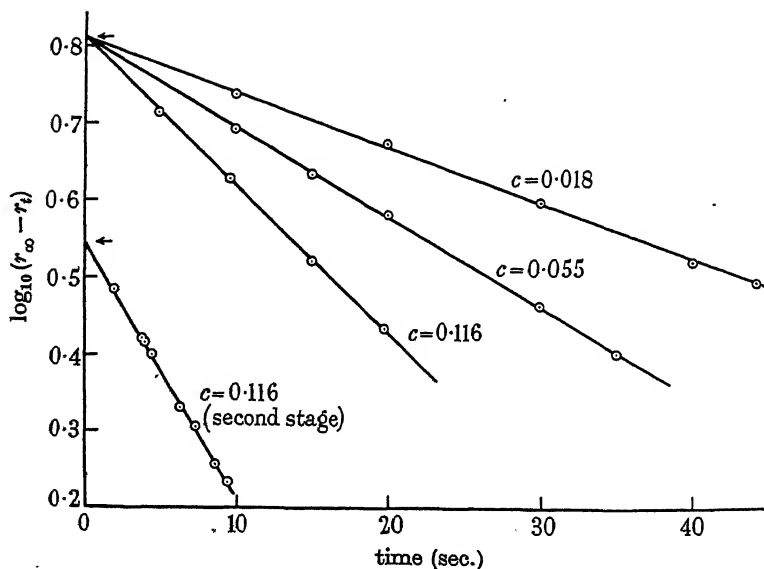


FIGURE 2

corrected for the acid produced and for the hydrolysis of bromine as described in earlier papers (Bell *et al.* 1948, p. 483). The values of k_2 are interpolated from the plot shown in figure 3.

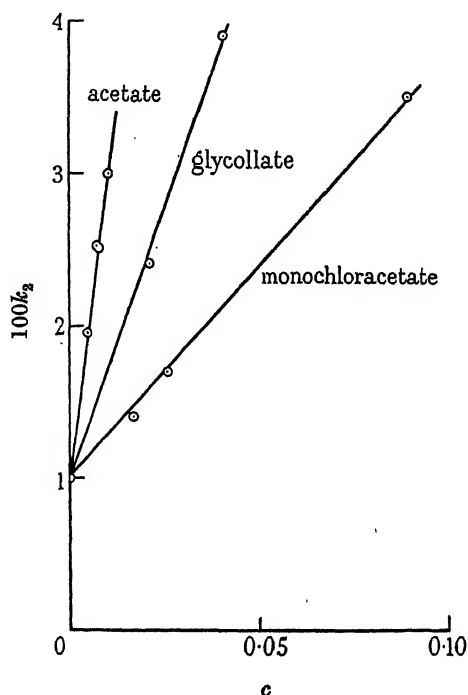


FIGURE 3

In the experiments with monochloracetate solutions recorded in table 1 a considerable proportion of the observed reaction is due to water catalysis, and the observed value of r is therefore intermediate between the value for water ($r_0 = 1.35$) and the value r_c for catalysis by monochloracetate. If K_2 and k_2 are the observed velocity constants for the bromination of bromoacetylacetone in water and in monochloracetate solution respectively, the value of the ratio r obtained by the above procedure is given by

$$r = k_2/k_1 = k_2 / \left\{ \frac{K_2}{r_0} + \frac{(k_2 - K_2)}{r_c} \right\}, \quad (5)$$

whence the value of r_c (given in the last column of table 3) can be obtained in terms of measured quantities.

The data for the other anions are similar, and only a summary will be given (table 2). The catalytic constants (k_b) for bromoacetylacetone were obtained from the plot of k_2 against c (see figure 3), and those for acetylacetone derived by using the values of r_c calculated as in table 1. To facilitate comparison with the results of Bell & Lidwell (1940) the catalytic constants are expressed in l./mole min., i.e. the velocity constants of table 1 have been multiplied by 2.303×60 . The catalytic constants for

bromoacetylacetone are probably not more accurate than $\pm 5\%$ because of the very high rates studied, and those for acetylacetone will be subject to somewhat greater errors in view of the indirect method used for deriving them.

TABLE 1. MONOCHLORACETATE CATALYSIS IN THE
BROMINATION OF ACETYLACETONE

c = anion concentration. buffer ratio (acid/anion) = 0.50.
 k_1, k_2 = first-order constants, sec.^{-1} , \log_{10} .
(a) calculated from r and k_2 . (b) from plot of equation (2).

c	$10^4 k_2$	value assumed for r	$10^4 k_1$			best value of r	r_c
			(a)	(b)	diff.		
0	100	1.2	83	80	- 3	1.3 ₅	—
		1.3	77	76	- 1		
		1.4	71	72	+ 1		
		1.5	67	70	+ 3		
0.018	150	1.3	115	106	- 9	1.5	1.9
		1.4	107	103	- 4		
		1.5	100	100	0		
0.055	250	1.3	193	167	-26	1.6	1.8
		1.4	179	172	- 7		
		1.5	167	163	- 4		
		1.6	156	155	- 1		
		1.7	147	152	+ 5		
0.116	430	1.6	268	269	+ 1	1.6	1.7
mean $r_c = 1.8$							

TABLE 2

catalyst	catalytic constant for		r_c
	acetylacetone	bromoacetylacetone	
water	1.03*	1.39*	1.3 ₅
chloroacetate	20	36.5	1.8
<i>m</i> -nitrobenzoate	45	195	4.3 ₅
glycollate	56	100	1.8
β -chloropropionate	77	225	2.9
phenylacetate	90	400	4.5
benzoate	96	280	2.9
acetate	160	270	1.7
β -phenylpropionate	180	780	4.2
propionate	190	370	1.9
trimethylacetate	220	720	3.3
isovalerate	250	600	2.3 ₅

* Velocity constant in slightly acid solution.

The bromination of acetylacetone has been previously studied by Bell & Lidwell (1940), but their results are less accurate than those obtained here, since they assumed that the second bromine atom is introduced much more quickly than the

first. In our present units their values for k_b were: water 1.37, monochloracetate 19, glycollate 47, acetate 154, trimethylacetate 254. Our present values differ somewhat from these, but do not involve any change in their general conclusions.

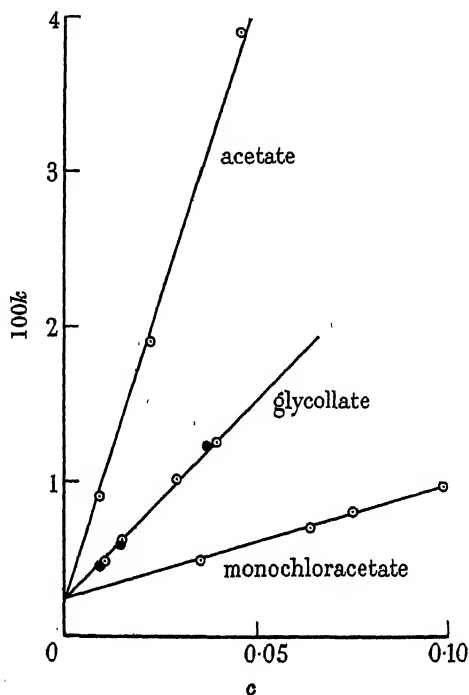


FIGURE 4

TABLE 3. CATALYTIC CONSTANTS FOR THE BROMINATION OF BENZOYLACETONE

(l./mole, min.)			
catalyst	k_b	catalyst	k_b
water	0.66*	phenylacetate	207
monochloracetate	22	acetate	222
glycollate	75	trimethylacetate	400
β -chloropropionate	141	isovalerate	420
<i>m</i> -nitrobenzoate	144	β -phenylpropionate	580
benzoate	146		

* Velocity constant in slightly acid solution.

In all the experiments with benzoylacetone the plot of $\log_{10}(r_{\infty} - r_t)$ against t for the overall reaction had the same slope as the corresponding plot for the second stage of the reaction. This is illustrated by figure 4 which shows the slope k as a function of anion concentration c for three catalysts. Most of the points represent the overall reaction, but for glycollate catalysis the solid circles refer to the second stage; similar results were obtained for the other catalysts, where the points for the two stages lie too closely together to be plotted separately.

From equation (1) this identity of slopes can arise only if $k_2 = \frac{1}{2}k_1$, i.e. $r = \frac{1}{2}$. This appears to be so within experimental error for all the catalysts studied, and we have therefore tabulated only the catalytic constants for benzoylacetone (table 3); those for bromobenzoylacetone are half as great.

MEASUREMENTS WITH ETHYL ACETOACETATE, ETHYL
 α -BROMOACETOACETATE AND OTHER SUBSTRATES

The bromination of ethyl acetoacetate has been studied by Pedersen (1933, 1934), who concluded from a laborious analysis of the rate of overall bromination that the α -bromoester was brominated about 12 times as fast as the original ester, this ratio varying little from one catalyst to another. He considered the possibility of bromination experiments with the α -bromoester, but stated that it was insufficiently stable for quantitative measurements. We have made further measurements on the rate of bromination of ethyl α -bromoacetoacetate, partly by preparing it *in situ*, and partly by using the pure ester; our results differ considerably from those of Pedersen.

Ethyl acetoacetate is a considerably weaker acid than acetylacetone or benzoylacetone, having a dissociation constant of 2×10^{-11} (Eidinoff 1945), and it therefore requires a very large excess of sodium hydroxide to convert it completely to the enolate ion. If bromination experiments are carried out with those solutions the initial titres are much less than those calculated for 100 % conversion to enol; this is due to the alkaline cleavage of the acetoacetic ester, and precludes the direct use of the method described in the last section for measuring the rate of bromination of the monobromoester. It is, however, possible to prepare sufficiently stable solutions with hydroxyl-ion concentration up to about 5×10^{-3} , in which the enol content is known either by calculation or by extrapolating the observed titres to zero time. If the fraction of enol thus found is E , the course of the bromination is given by equation (1) as

$$(r_\infty - r_t)/r_\infty = \frac{(1-E)(2r-1)}{2(r-1)} 10^{-k_1 t} - \frac{(1-rE)}{2(r-1)} 10^{-k_2 t}, \quad (6)$$

where r_t is the titre after time t . For the bromination of acetoacetic ester to which no alkali had been added, the corresponding equation is

$$(r_\infty - r'_t)/r_\infty = \frac{2r-1}{2(r-1)} 10^{-k_1 t} - \frac{1}{2(r-1)} 10^{-k_2 t}, \quad (7)$$

neglecting the very small proportion of enol (0.4 %) in the aqueous solution. Combining equations (6) and (7) we find

$$\{(r_\infty - r_t) - (1-E)(r_\infty - r'_t)\}/r_\infty = \frac{1-(r-1)E}{2(r-1)} 10^{-k_2 t}, \quad (8)$$

so that a plot of $\log_{10}\{(r_\infty - r_t) - (1-E)(r_\infty - r'_t)\}$ against t should be a straight line of slope k_2 .

Two typical experiments used solutions of the following compositions:

	initial ester concentration	concentration of NaOH added	% enolate	
			calc.	obs.
1	3.56×10^{-4}	8.1×10^{-4}	54	50
2	3.65×10^{-4}	56.0×10^{-4}	90	88

The acetoacetic ester was purified by two distillations under reduced pressure, and the technique used was exactly the same as for benzoylacetone and acetylacetone. The bromine solution contained enough acid to neutralize all the alkali present, so that the value of k_2 represents rate of bromination of α -bromoacetoacetic ester in the absence of catalysts other than the water molecule. Some cleavage of acetoacetic ester or its bromination products takes place in the later stages of the bromination, and the titres for long reaction times fall short of the theoretical r_∞ value. However, this complication does not affect the first 2 to 3 min. of the reaction, and figure 5 shows the plots of $\log_{10}\{(r_\infty - r_t) - (1 - E)(r_\infty - r'_t)\}$ against t for the two experiments listed above. The plots are linear and parallel within the experimental error. In further confirmation of the value of k_2 we have carried out experiments with pure α -bromoacetoacetic ester, prepared by the method of Brühl (1903) and purified by distillation under reduced pressure (b.p. 97 to 98°/8 mm.). An aqueous solution prepared from freshly distilled ester liberated the theoretical amount of iodine from potassium iodide, and the titre remained constant for some hours, after which it fell slowly, probably because of the transformation of α -bromoester into γ -bromoester (see Macbeth 1923; Kharasch, Sternfeld & Mayo 1937). The bromination experiments were carried out with freshly prepared solutions and lasted only a few minutes. The plot marked 3 in figure 5 refers to such an experiment in slightly acid solution; it is parallel to the other two plots within experimental error.

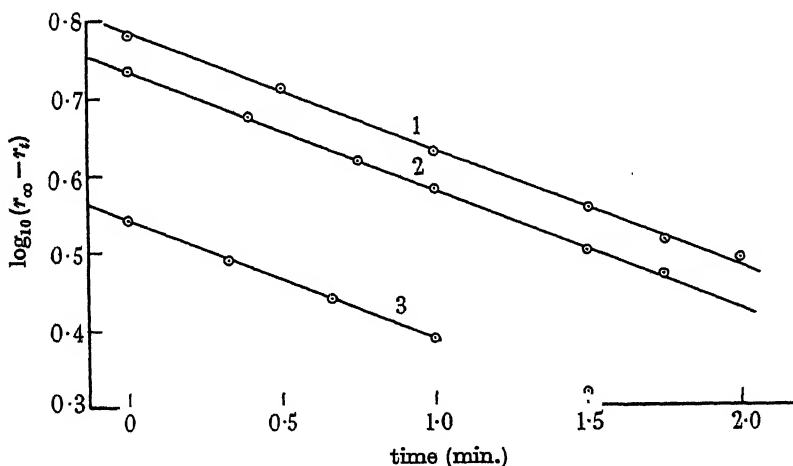


FIGURE 5

The slope of the lines in figure 5 corresponds to $k_2 = 0.15$ (min.^{-1} , \log_{10}) compared with the value of 0.38 given by Pedersen (1933). We are unable to account for this discrepancy, especially since later work by the same author using a different experimental technique (Pedersen 1948) also gives $k_2/k_1 \sim 11$. However, the method of calculation used by Pedersen is insensitive to variations in k_2 , and may have been affected by a side-reaction taking place to a small extent. In any case the value obtained for k_1 (corresponding to the introduction of the first bromine atom into acetoacetic ester) will depend very little on whether $k_2/k_1 \sim 5$ (as we have found) or

$k_2/k_1 \sim 10$ (as given by Pedersen); in either case a direct plot of $\log_{10}(r_\infty - r_t)$ against t for the overall reaction will give k_1 without serious error. Thus for the bromination in acid solution we find by this method $k_1 = 0.030$, while Pedersen gives $k_1 = 0.0312$.

In investigating the effect of anion catalysts we have therefore not attempted to evaluate k_2 , but have determined k_1 (and the corresponding catalytic constants) from logarithmic plots of the overall bromination rate. The results obtained are given in table 4.

TABLE 4. CATALYTIC CONSTANTS FOR THE BROMINATION OF ETHYL ACETOACETATE

(l./mole, min.)			
catalyst	k_b	catalyst	k_b
water	0.072*	propionate	32
monochloracetate	2.28†	β -phenylpropionate	34
<i>m</i> -nitrobenzoate	6.5	isobutyrate	39
glycollate	8.0	isovalerate	41
benzoate	15	trimethylacetate	43
phenylacetate	20	<i>n</i> -valerate	44
acetate	29†		

* Velocity constant in slightly acid solution.

† Values due to Pedersen (1934).

For comparison with α -bromoacetoacetic ester some experiments were also made with bromomalonic ester. Previous work (Bell *et al.* 1946) had indicated that this ester was brominated considerably more rapidly than malonic ester itself, but no quantitative data were available. Bromomalonic ester was prepared as described in *Organic Syntheses* (Collective vol. 1, p. 240) and had b.p. 121 to 122° C/14 mm. Bromination experiments were carried out in slightly acid solutions containing about 4×10^{-4} M ester, using the technique described by Bell *et al.* (1946). The plots of $\log_{10}(r_\infty - r_t)$ against t were linear over approximately the first half of the reaction, and two experiments gave for the first-order constant (in min.⁻¹, \log_{10}) $10^3 k_2 = 8.0$, and 7.7; mean $k_2 = 7.9 \times 10^{-3}$. This represents the velocity of the water-catalyzed reaction, and the corresponding velocity for malonic ester itself is 6.3×10^{-4} .

DISCUSSION

In previous work on base-catalyzed reactions relations have been found to exist between the catalytic constant of an anion, k_b , and its basic strength, conveniently measured by $1/K_A$, where K_A is the dissociation constant of the corresponding acid. For halogenation reactions these relations have only been tested for a very limited number of catalysts, and it is of interest to see how far they are valid for the more extensive range of catalysts studied here. Figure 6 shows a plot of $\log_{10} k_b$ against $\log_{10} K_A$ for the bromination of acetoacetic ester and acetylacetone (see tables 2 and 4), and it will be seen that the relation is obeyed reasonably well in each case. This is not the case in the bromination of bromoacetylacetone and benzoylacetone (see figure 7), where there is a general trend towards an increase of k_b with decreasing K_A , but no quantitative relation.

In seeking an explanation of these divergencies, it is noteworthy that the points for catalysis by acetate, glycollate and monochloracetate ions lie accurately on a straight line for each of the substrates studied, while the remaining catalysts, which contain large groups (alkyl, aryl or bromo-groups) attached to the carbonyl, all give reaction velocities greater than those interpolated from this line. These positive

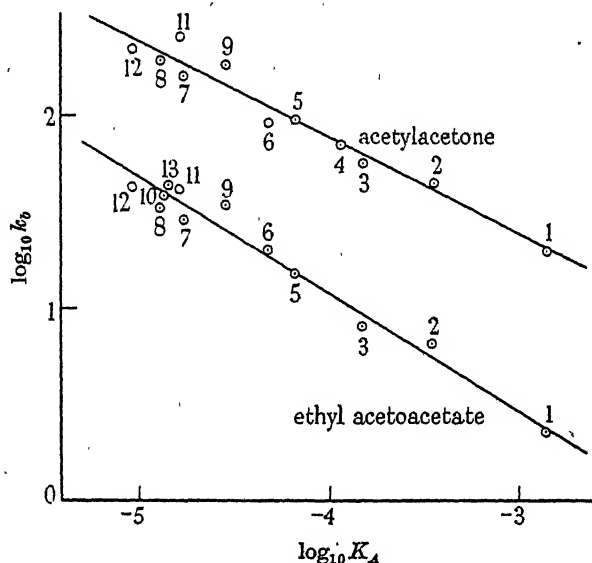


FIGURE 6

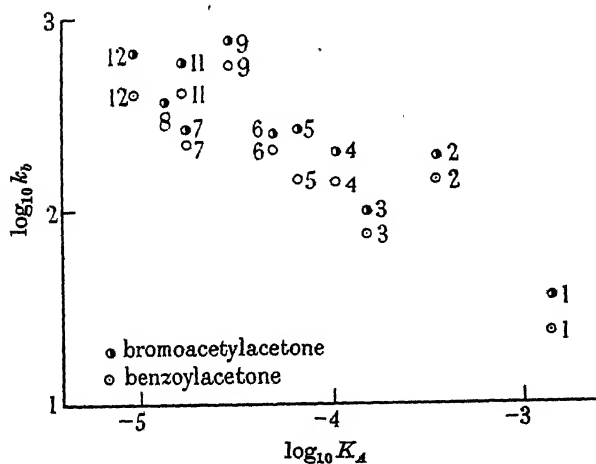


FIGURE 7

Key to figures 6 and 7.

- | | | |
|------------------------------|------------------------------|------------------------|
| 1. monochloracetate | 6. phenylacetate | 11. isovalerate |
| 2. <i>m</i> -nitrobenzoate | 7. acetate | 12. trimethylacetate |
| 3. glycollate | 8. propionate | 13. <i>n</i> -valerate |
| 4. β -chloropropionate | 9. β -phenylpropionate | |
| 5. benzoate | 10. isobutyrate | |

deviations are detectable for acetoacetic ester and acetylacetone, and become large (up to nearly 300 %) for bromoacetylacetone and benzoylacetone, which also contain large groups. It therefore seems reasonable to attribute such deviations to the proximity of a large group in the catalyst to a similar group in the substrate, which will affect the transition state of the reaction, but will not contribute anything to the dissociation constant which is used as a measure of the basic strength of the catalyst. There will be some direct interaction between the two groups, but a more important factor is probably the necessity of making a cavity in the solvent, thereby doing away with some of the interactions between the water molecules. When the two groups are close together they will cause the separation of fewer water molecules than when they are apart, and this factor will tend to make the transition state more stable. The order of magnitude of such an effect can be estimated from the work of Butler (1937) on the solubilities of homologous series of molecules in water: he found that the addition of each CH_2 group caused on the average an increase of 160 cal./mole in the free energy of solution, and of 1570 cal./mole in the heat of solution. An effect of this kind would therefore be quite adequate to account for the increases of reaction velocity referred to above, the largest of which corresponds to a decrease of 800 cal./mole in the free energy of the transition state.

It seems justifiable to consider that catalysis by the acetate, glycolate and chloracetate ions does not involve effects of this kind, since they do not contain large groups, and they conform accurately to an equation of the type

$$k_b = G(1/K_A)^\alpha \quad (9)$$

for each reaction studied. We have therefore used the results for these three ions (and for water catalysis) in making a comparison between the various ketonic substances whose halogenation has been studied. Previous work has shown that there is a correlation between the value of α in equation (9) and the reactivity of the substrate. The latter is conveniently measured by R , which is a measure of the rate of ionization of the ketone in presence of unit concentration of a catalyst with $K_A = 10^{-4}$, corrected for statistical differences between the different ketones. In previous papers k_b has not always been given in the same units and we have, therefore collected all the available data in table 5, in which R is in l./mole, min. We have also modified the method of making a statistical correction for the presence of more than one ionizable hydrogen in the substrate molecule. In previous papers (Bell & Lidwell 1940; Bell 1943) several hydrogen atoms on the same carbon were reckoned as one hydrogen, while in table 5 they are counted separately; e.g. the velocity for acetone has been divided by 6 rather than by 2. We have made this modification because bromobenzoylacetone has been found to brominate only half as fast as benzoylacetone itself, and it seems unlikely that the introduction of a bromine atom should decrease the tendency to ionize. The choice of statistical correction does not affect any of the main conclusions drawn. The values of α and R for α -bromoacetoacetic ester have been omitted from the table, since the work described in the preceding section makes it probable that the values given by Pedersen (1933) for this substance are in error, and our present values for acetylacetone have been substituted for the earlier values of Bell & Lidwell. It will be seen that the decrease

in α with increasing reactivity, originally suggested on the basis of less complete data, is fully borne out by the fuller and more accurate results now available. The only marked exception to this regularity is now acetoacetic acid, and the value of α for this substance is not known with certainty, being based on two catalysts only, and on extrapolation from 18 to 25° C. On the other hand, the explanation originally advanced to explain the decrease in α (Lidwell & Bell 1940) is probably not wholly

TABLE 5

$R = \text{catalytic constant in l./mole, min.}$ $v_0 = \text{water velocity in min.}^{-1}$				
} statistically corrected				
substrate	$\log_{10} R$	α	$\log_{10} v_0$	source
CH_3COCH_3	-6.78	0.88	-8.33	<i>a</i>
$\text{CH}_3\text{COC}_6\text{H}_5$	-6.07	—	—	<i>b</i>
$\text{CH}_3\text{COCH}_2\text{CH}_2\text{COCH}_3$	-5.46	0.89	-7.45	<i>a</i>
$\text{CH}_3\text{COCH}_2\text{Cl}$	-3.51	0.82	-5.79	<i>a</i>
$\text{CH}_3\text{COCH}_2\text{Br}$	-3.25	0.82	-5.10	<i>a</i>
$\text{CH}_3\text{COCHCl}_2$	-2.00	0.82	-4.36	<i>a</i>
$\text{CH}_2(\text{COOC}_2\text{H}_5)_2$	-0.76	0.79	-3.14	<i>a</i>
$\text{CH}_3\text{COCH}_2\text{COOC}_2\text{H}_5$	+0.72	0.59	-1.44	<i>a</i>
$\text{CO}(\text{CH}_2)_3\text{CHCOOC}_2\text{H}_5$	+1.18	0.58	-1.85	<i>d</i>
$\text{CHBr}(\text{COOC}_2\text{H}_5)_2$	—	—	-1.74	<i>e</i>
$\text{CH}_3\text{COCH}_2\text{COC}_6\text{H}_5$	+1.33	0.52	-0.48	<i>e</i>
$\text{CH}_3\text{COCHBrCOC}_6\text{H}_5$	+1.33	0.52	-0.48	<i>e</i>
$\text{CH}_3\text{COCHBrCOOC}_2\text{H}_5$	—	—	-0.46	<i>e</i>
$\text{CH}_3\text{COCH}_2\text{COCH}_3$	+1.54	0.48	-0.29	<i>e</i>
$\text{CH}_3\text{COCHBrCOCH}_3$	+2.04	0.42	+0.11	<i>e</i>
$\text{CH}_3\text{COCH}_2\text{COOH}$	+2.27	0.48	+0.30	<i>a</i>

References: *a*, Bell & Lidwell 1940; *b*, Bell 1943; *c*, Bell *et al.* 1946; *d*, Bell *et al.* 1948; *e*, present paper.

correct. This explanation assumed that the potential energy curves for proton-transfer had the same shape over the whole range of ketones concerned, and that the varying value of α corresponded to the change of slope on moving up and down this curve. It was difficult to account on this basis for the large changes in α found experimentally and more recently (Bell & Higginson 1949) both experimental and theoretical evidence has been presented to show that the shape of the potential energy curves will depend on the extent to which a change of electronic structure takes place on the transfer of a proton. This will account for the low values of α found for the β -diketones and ketonic esters in table 5, for which the charge in the anion is spread fairly evenly over two oxygen atoms. It is of course this same spread of charge (or mesomerism) which makes these ketones relatively strong acids and gives them high rates of ionization.

Other points of structural interest emerge from table 5, notably the effect on the ionization velocity of substituting one hydrogen by bromine in the active >CH_2 group. The data are most complete for the water reaction, though essentially the same picture is given by the results for anion catalysis. The relevant figures are given in table 6. The effect of bromine substitution decreases in a striking manner as the

reactivity of the ketone increases, and is relatively small in the β -diketones. This can again be related to the spread of charge in the anions of the latter. If the bromine can increase the stability of the anion by the attraction between the negative charge and the positive end of the C^+-Br^- dipole, then this stabilization will be diminished when the charge is removed almost completely from the carbon by being spread more or less evenly over two oxygens.

TABLE 6. EFFECT OF BROMINE SUBSTITUTION

$r' = (v_0 \text{ for bromoketone})/(v_0 \text{ for unsubstituted ketone})$, using values corrected statistically.

ketone	v_0	r'
acetone	4.5×10^{-3}	1700
diethylmalonate	7×10^{-4}	25
acetoacetic ester	3.6×10^{-2}	10
benzoylacetone	3.3×10^{-1}	1
acetylacetone	5×10^{-1}	2.7

Our thanks are due to the International Federation of University Women for a grant to one of us (E.M.).

REFERENCES

- Bell, R. P. 1943 *Trans. Faraday Soc.* **39**, 253.
 Bell, R. P., Everett, D. H. & Longuet-Higgins, H. C. 1946 *Proc. Roy. Soc. A*, **186**, 443.
 Bell, R. P. & Higginson, W. C. E. 1949 *Proc. Roy. Soc. A*, **197**, 141.
 Bell, R. P., Lidwell, O. M. & Vaughan-Jackson, M. W. 1936 *J. Chem. Soc.* p. 1794.
 Bell, R. P. & Lidwell, O. M. 1940 *Proc. Roy. Soc. A*, **176**, 88.
 Bell, R. P., Smith, R. D. & Woodward, L. A. 1948 *Proc. Roy. Soc. A*, **192**, 479.
 Brühl, J. W. 1903 *Ber. dtsch. chem. Ges.* **36**, 1731.
 Butler, J. A. V. 1937 *Trans. Faraday Soc.* **33**, 229.
 Eidinoff, M. L. 1945 *J. Amer. Chem. Soc.* **67**, 2073.
 Kharasch, M. S., Sternfeld, E. & Mayo, F. R. 1937 *J. Amer. Chem. Soc.* **59**, 1655.
 Lidwell, O. M. & Bell, R. P. 1940 *Proc. Roy. Soc. A*, **176**, 114.
 Macbeth, A. K. 1923 *J. Chem. Soc.* p. 1122.
 Peat, S., Bourne, E. J. & Thrower, R. D. 1947 *Nature*, **159**, 811.
 Pedersen, K. J. 1933 *J. Phys. Chem.* **37**, 751.
 Pedersen, K. J. 1934 *J. Phys. Chem.* **38**, 601, 999.
 Pedersen, K. J. 1948 *Acta Chem. Scand.* **2**, 252.

The catalytic hydrogenation of methyl elaeostearate, and of mixtures of elaeostearic with other polyethenoid long-chain esters

By T. P. HILDITCH, F.R.S., AND S. P. PATHAK

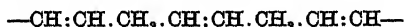
Department of Industrial Chemistry, University of Liverpool

(Received 15 March 1949)

The hydrogenation (using Raney nickel catalyst at 110 and 170° C) of the conjugated triene system present in methyl elaeostearate (octadeca-9, 11, 13-trienoate) has been compared with that of the pentadiene systems $\text{—CH:CH.CH}_2\text{.CH:CH—}$ present in methyl linolenate and methyl linoleate.

The course of the hydrogenation action in methyl elaeostearate itself has been followed spectrophotometrically, and it has been established that, for the most part, the primary action is the simultaneous addition of two molecules of hydrogen to one molecule of elaeostearate and formation of very large proportions of octadeca-11-enoic esters. Production of stearate does not occur to any marked extent until over 80 % of elaeostearate has been converted to mono-ethenoid ester.

The selectivity of the hydrogenation process has been studied in the cases of equimolecular binary mixtures of methyl elaeostearate with, respectively, methyl oleate, linoleate or linolenate. In a mixture of methyl elaeostearate and oleate, hydrogenation of the latter does not set in until nearly all the elaeostearate has been converted to octadecenoate. With a mixture of methyl elaeostearate and linoleate, the diene ester remains almost unattacked until 50 % or more of the conjugated triene ester present has been converted into mono-ethenoid ester, after which it also commences to be transformed into mono-ethenoid esters. In a mixture of methyl elaeostearate and linolenate, both esters undergo hydrogenation from the outset, but conjugated triene ester disappears about twice as rapidly as the non-conjugated isomer. The activity to hydrogen of the double pentadiene system



is thus comparable with that of the conjugated triene group in the elaeostearate molecule.

The hydrogenation of methyl linoleate in presence of Raney nickel either at 110 or 170° C is extremely selective, no methyl stearate being produced until over 90 % of the linoleate has been transformed into octadecenoates.

INTRODUCTION

The phenomenon of selective hydrogenation—the preferential attachment of hydrogen in presence of a catalyst to one of two or more centres of unsaturation—has been recognized for many years, both in compounds belonging to the terpene series and in the long-chain unsaturated acyl groups present in the mixed glycerides of the fats (for a summary of the earlier observations, see Armstrong & Hilditch 1925). In the aliphatic series, pronounced selectivity of the hydrogenation process was first observed in certain vegetable fatty oils (H. K. Moore, Richter & van Arsdel 1917; C. W. Moore & Hilditch 1923), in which the linoleic glycerides pass almost wholly to the mono-ethenoid state before any saturated stearic glycerides are formed. Richardson, Knuth & Milligan (1925) reported that the multi-ethenoid glycerides present in whale and other marine animal oils behave differently, in that when saturated glycerides have been produced in quantity, the remaining unsaturated

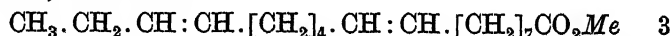
glycerides still contain notable amounts of di-ethenoid compounds; this was later confirmed by Hilditch & Terleski (1937) and Harper & Hilditch (1937).

This seeming difference in behaviour remained unexplained until a detailed study of the hydrogenation of the triene methyl linolenate (octadeca-9, 12, 15-trienoate) by Bailey & Fisher (1946) showed that the primary products are octadeca-9, 15-dienoate with about an equal amount of a mixture of octadeca-9, 12- and -12, 15-dienoates, and that to a minor extent two molecules of hydrogen are added simultaneously to linolenate with direct formation of mono-ethenoid esters. Further, these workers gave the relative reactivities of the different unsaturated esters as approximately:

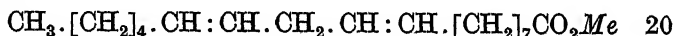
methyl oleate



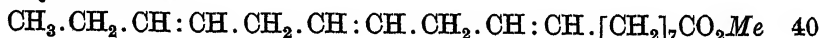
methyl octadeca-9, 15-dienoate



methyl linoleate

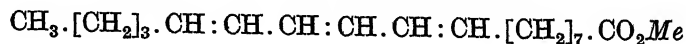


methyl linolenate



These results lead unequivocally to the conclusion (Hilditch 1946) that selective hydrogenation in long-chain unsaturated compounds is essentially that of the specific pentadiene system $-\text{CH} : \text{CH} \cdot \text{CH}_2 \cdot \text{CH} : \text{CH}-$ in which a 'reactive' methylene group occurs between two double bonds. The apparent persistence of di-ethenoid unsaturation during hydrogenation of a compound which initially contained a multi (4, 5, or 6)-ethenoid system (as in the marine animal oils) is now explicable, since primary attack of hydrogen at one or more of the pentadiene groupings originally present must result in the production of di- or even tri-ethenoid derivatives, in which the remaining double bonds are separated by four or more methylene groups, and which consequently react to catalytic hydrogenation little more rapidly than a mono-ethenoid compound.

The specific properties of the pentadiene grouping towards hydrogen in presence of a catalyst, thus made clear, rendered it of interest to compare the behaviour of the corresponding conjugated triene (elaeostearic) system



with that of methyl linolenate, linoleate and oleate. The hydrogenation of methyl elaeostearate itself evidently required further study, since the course of the action had been reported differently by earlier workers. Boeseken & Hoogland (1927) and Boeseken, van Krimpen & Blanken (1930) stated that reduction of elaeostearic glycerides (tung oil) proceeds in accordance with the Thiele mechanism, molecules of hydrogen being successively added to produce first a conjugated octadeca-10, 12-dienoate and then octadec-11-enoate, before any saturated derivatives appear. Steger, van Loon & van Vlimmeren (1944), however, found that conjugated unsaturation disappeared steadily from the commencement of hydrogenation of the oil, i.e. no conjugated di-ethenoid reduction products are formed. They confirmed the

observation that no saturated compounds are produced until the reaction product consists practically entirely of mono-ethenoid compounds, but considered that in the first phase of the hydrogenation non-conjugated diene derivatives are produced. Groot, Kentie & Knol (1947) later recommended the partial hydrogenation of tung oil as a means of obtaining almost pure octadec-11-enoic acid.

METHODS

The hydrogenations were carried out (on 50 to 120 g. of esters) in a three-necked round-bottomed flask fitted with a mechanical stirrer (1300 r.p.m.) and maintained at either 110 or 170°C by an oil bath. Before commencing an experiment the apparatus was evacuated and then filled with hydrogen; during the actual hydrogenation the volume of hydrogen absorbed was followed approximately by delivering it to the reaction flask from a graduated gas container. Samples of the hydrogenated product were withdrawn for examination at successive intervals corresponding approximately with the absorption of about one-sixth of the volume of hydrogen necessary for complete conversion to saturated ester (stearate).

The catalyst used was Raney nickel, which was prepared from powdered nickel-aluminium alloy as described by Pavlic & Adkins (1946) and kept under alcohol until required for use. The concentration of catalyst employed was approximately 2% of the esters.

Preparation of the unsaturated esters

The requisite fatty acids were (except in the case of methyl linolenate) prepared from natural sources by means of low-temperature crystallization from appropriate solvents. In no case were they of 100% purity, but consisted of materials containing high concentrations (usually over 90%) of the desired compound accompanied by minor proportions of known constituents which did not interfere with the object of the experiments. In view of the liability of long-chain polyethenoid compounds to suffer change during chemical handling, it was decided that this course was to be preferred to the more lengthy procedure necessary in order to obtain a completely individual ester.

Methyl elaeostearate

Since elaeostearic acid tends to polymerize during handling and especially in presence of reagents such as sulphuric acid, and since methyl elaeostearate also commences to undergo polymerization during distillation at 0.2 mm. pressure through a fractionation column, the method finally adopted was to esterify the mixed acids from tung oil (isolated therefrom by rapid hydrolysis with a very small excess of alkali) with methyl alcohol at room temperature in presence of about 0.5% of anhydrous hydrogen chloride. Esterification was complete after about 24 hr., and the elaeostearic content (77%) of the mixed esters (after removal of any small amounts of free acids present) was found to be the same as that of the tung oil from which they had been prepared. The mixed methyl esters (365 g.) were then crystallized from 10% solution in acetone at -40°C, when 212 g. remained in solution; these esters, when further crystallized from 10% solution in acetone at -60°C,

deposited 103 g. of esters which contained 94 % of methyl elaeostearate (with 5 % of oleate and 1 % of saturated esters). This (and analogous batches of similarly prepared methyl elaeostearate) was employed in the present hydrogenation studies.

Methyl oleate

Neat's foot oil is a convenient source of oleic acid, since the latter (64 %) is accompanied in this oil by relatively small proportions of linoleic (3 %) and other polyethenoid acids (2 %). The bulk of the saturated acids present in the mixed fatty acids from the oil were removed in the form of lead salts insoluble in alcohol. The acids from the soluble lead salts (e.g. 217 g., iodine value 91.6) were first crystallized from 10 % solution in acetone at -40°C , when 121 g. (iodine value 86.8) were deposited and recrystallized from acetone at the same temperature, when 102 g. (iodine value 83.5) were deposited. These acids were converted into methyl esters which were fractionally distilled in order to remove methyl esters of lower acids (mainly palmitic and hexadecenoic); 66 g. of methyl oleate (iodine value 84.0; equivalent 295.0; calc. iodine value 85.8, equivalent 296.0) were obtained and used in the present experiments. Linoleic ester was found to be absent from this ester, which contained 98 % of methyl oleate and 2 % of methyl stearate.

Methyl linoleate

Sunflower-seed oil with a content of 70 % of linoleic acid was used as the source of concentrates of methyl linoleate accompanied by methyl oleate. When the mixed acids of sunflower-seed oil are crystallized from acetone at -60°C , the material left in solution (iodine value 165 or above) is rich in linoleic acid (iodine value 181.4). Further crystallization of the methyl esters of this concentrate from acetone at -65 or -70°C usually leads to the deposition of esters somewhat richer in linoleate (e.g. iodine value 162, calc. 172.8). Esters of similar iodine value were finally distilled at 0.2 mm. through a fractionation column, and the fractions of highest iodine value were used in the hydrogenation experiments. In all, 163 g. of material of iodine value 163.3 was obtained (methyl linoleate 88 %, methyl oleate 12 %).

Methyl linolenate

It is not yet possible to separate linolenic acid from linoleic acid sufficiently for the present purpose by crystallization at low temperatures. Linolenic acid was therefore prepared by debromination of hexabromostearic acid (m.p. 180°C) with zinc and pyridine (Kaufmann & Mestern 1936) and converted into crude methyl linolenate, which when distilled at 0.2 mm. pressure yielded methyl linolenate (iodine value 258, calc. 262.0).

Determination of the composition of the products of hydrogenation

(i) *Spectrophotometric methods.* Conjugated triene systems in long-chain acids absorb ultra-violet light in the region of 265 to 275 $\text{m}\mu$ with a well-defined absorption band head at 268 $\text{m}\mu$ (extinction coefficient $E_{1\text{cm}}^{1\%}$ 1780); analogous conjugated diene acids give an absorption band in the ultra-violet spectrum at 234 $\text{m}\mu$ ($E_{1\text{cm}}^{1\%}$ 1200). Spectrophotometric measurement of extinction coefficients at these wave-lengths

thus leads directly to the determination of the proportions of conjugated triene and diene compounds present.

The spectrophotometric method can also be applied to the determination of linolenic and linoleic compounds (Hilditch, Morton & Riley 1945). These, although non-conjugated, contain the system $-\text{CH}:\text{CH}.\text{CH}_2\text{CH}:\text{CH}-$, which, on treatment with a high concentration of alkali hydroxide in an appropriate solvent at about 180°C under standardized conditions, undergoes rearrangement to conjugated isomers (diene from linoleic compounds, diene and also triene from linolenic compounds). In the present work the analytical conditions recommended by Hilditch *et al.* (1945) were observed, namely, for linoleic compounds, isomerization with alkali in glycol solution at 180°C for 60 min. and for linolenic compounds similar isomerization at 170°C for 15 min. The amount of elaeostearic, linolenic, linoleic and any conjugated diene compounds present in a mixture containing some or all of these unsaturated compounds can thus be determined from the spectrophotometric observations described, due allowance being made, when determining linolenic acid, for the contribution to the extinction coefficient observed at $268\text{ m}\mu$ after alkali-isomerization due to elaeostearic acid; whilst, when determining linoleic acid, similar deductions from the observed extinction coefficient at $234\text{ m}\mu$ after alkali-isomerization must be made for the contributions due to elaeostearic, linolenic or any conjugated diene acid.

The reference values employed for the extinction coefficients $E_{1\text{ cm}}^{1\%}$, for the respective pure acids, whether as such or after alkali-isomerization under the specified conditions, were as follows:

	unisomerized		after alkali-isomerization	
	$268\text{ m}\mu$	$234\text{ m}\mu$	$268\text{ m}\mu$	$234\text{ m}\mu$
elaestearic acid	1780	208	1690	197
linolenic acid	—	—	532	569
conjugated diene acid	—	1200	—	1140
linoleic acid	—	—	—	906

(ii) *Determination of total unsaturation (iodine value).* The total amount of ethenoid unsaturation in long-chain aliphatic compounds can be determined by various methods which depend upon the quantitative addition of iodine mono-halides to the double bonds when the necessary conditions are observed. With non-conjugated unsaturated compounds quantitative addition takes place readily in the course of 30 min. in presence of a large excess of iodine monochloride in glacial acetic acid solution (Wijs 1898), but conjugated double bonds react only incompletely with the Wijs solution. When these are present, however, a reasonably accurate measure of the total unsaturation is obtained if the bromine vapour absorption method of Toms (1928) is employed, the result being expressed for convenience in terms of 'iodine value', i.e. the amount of halogen (calculated as iodine) which has combined with 100 parts of the unsaturated compound, or mixture of compounds. Whilst the accuracy of determination of unsaturation in mixtures containing conjugated unsaturated long-chain acids is not equal to that obtainable when only non-conjugated derivatives are present, the Toms method has proved sufficiently reliable

in the present investigation to enable us to follow the course of hydrogenation of methyl elaeostearate and of mixtures of this ester with analogous non-conjugated esters.

In general in a mixture of unsaturated long-chain esters, in which elaeostearate, linolenate and linoleate have been determined spectrophotometrically (as in (i) above), the proportion of oleic or other mono-ethenoid compounds then follows from the difference between the observed (Toms) iodine value and the sum of the increments of iodine value due to the observed amounts of the polyethenoid esters. Any saturated compounds present are then determined by difference.

(iii) *Determination of saturated acids in partially hydrogenated mixtures of esters.* In the special case of hydrogenation of methyl elaeostearate, or mixtures of this with other unsaturated esters, it was desired to determine the proportions of mono- and di-ethenoid esters present at progressive stages of the hydrogenation. This could be effected if the mean iodine value of the mono- and di-ethenoid esters could be calculated. For this purpose it was necessary separately to determine the proportion of saturated esters present, and the procedure of Bertram (1925) was employed. This method, although not of the highest order of analytical accuracy, gives results accurate to within about 1 %. The proportions of triene esters (elaestearate or linolenate) having been determined spectrophotometrically, and the proportion of saturated esters having been determined, the approximate iodine value of the remaining components of the system (mono- or di-ethenoid C_{18} esters) can be calculated from the iodine value (Toms) of the original mixture, whence the general proportions of the mono- and of di-ethenoid components can be arithmetically calculated.

RESULTS AND DISCUSSION

Hydrogenation of methyl linoleate with Raney nickel catalyst

The behaviour of methyl linoleate with the partly dispersed form of catalyst known as Raney nickel has not previously been recorded, and it seemed well to ascertain whether the selectivity observed with other forms of catalytic nickel was also shown in presence of Raney nickel, especially at the lower (110°C) of the two temperatures employed throughout this work. As the data in table 1 show, the hydrogenation was

TABLE 1. HYDROGENATION OF METHYL LINOLEATE

iodine value (Wijs)	linoleate (%)	mono-ethenoid esters (%)	saturated (%)
hydrogenation at 110°C			
163.4	88.3	11.7	—
133.1	55.4	44.0	0.6
101.7	18.4	81.4	0.2
78.7	1.7	88.1	10.2
hydrogenation at 170°C			
163.4	88.3	11.7	—
138.2	61.5	37.2	1.3
94.1	11.2	87.1	1.7
75.3	0.6	86.5	12.9

extremely selective both at 110 and 170° C, production of methyl stearate being negligible until 90 % or more of the linoleate had passed to the mono-ethenoid condition. The first lines of figures in both parts of table 1 refer to the original concentrate of methyl linoleate (cf. p. 326) used in the hydrogenations, succeeding lines referring to samples withdrawn as the hydrogenations proceeded. In this series of experiments the composition of the ester mixtures was determined from the iodine value and the extinction coefficient $E_{1\text{cm}}^{1\%}$ at 234 m μ after alkali-isomerization (for details see table 6).

Hydrogenation of methyl elaeostearate

A concentrate of methyl elaeostearate, prepared as described (p. 325) and containing 93 % of methyl elaeostearate with minor proportions of oleate and (possibly) linoleate, was hydrogenated with Raney nickel catalyst at 110 and at 170° C with the results summarized in table 2. In table 2 the data for elaeostearate were determined spectrophotometrically, saturated acids were determined by the Bertram method (1925), and the figures for mono- and di-ethenoid compounds were based upon the calculated iodine value of this group (i.e. after correcting the iodine value of the ester mixtures for the observed proportions of elaeostearic and saturated compounds).

TABLE 2. HYDROGENATION OF METHYL ELAEOSTEARATE

iodine value (Toms)	elaestearate (spectrographic) (% found)	saturated (Bertram) (% found)	di-ethenoid (from residual iodine value) (% calc.)	mono-ethenoid (from residual iodine value) (% calc.)
hydrogenation at 110° C				
251.5	93	—	3	4
218.0	86	4	—	10
199.1	73	7	—	20
169.2	47	10	14	29
128.0	27	13	7	53
76.3	6	22	—	72
50.5	—	50	9	41
hydrogenation at 170° C				
251.5	93	—	3	4
219.0	81	4	—	15
203.0	67	5	4	24
183.3	40	6	37	17
113.3	16	9	9	66
68.7	—	30	3	67

Rigidly accurate determination of the iodine value of esters containing high proportions of elaeostearate is difficult to accomplish (even by the bromine vapour absorption procedure), whilst the same applies in perhaps lesser degree to the quantitative determination of the saturated acids present. Moreover, the greater part of the (calculated) iodine value of the more unsaturated products is contributed by elaeostearate, so that the residual iodine value (which represents the mono- and di-ethenoid components of the mixture) is relatively small by comparison with the

total iodine value observed by bromine vapour absorption. Consequently, no high degree of accuracy is claimed for the calculated proportions of mono- and di-ethenoid esters, although these figures serve to point unequivocally to the course which the hydrogenation has for the most part followed.

The data in table 2 show that the proportion of saturated esters increases but little (especially at 170° C) until practically all the elaeostearate has disappeared. At the same time, the calculated figures in the two final columns indicate that (subject to one or two exceptions referred to below) little or no diethenoid esters are produced, the first stage of hydrogenation of the elaeostearate being very largely the simultaneous addition of four atoms of hydrogen with production of a molecule of a mono-ethenoid ester. The absorption spectra in the range 230 to 300 $m\mu$ of the original ester and of its four products of partial hydrogenation at 110° C with iodine values 218.0, 199.1, 169.2 and 128.0 (figure 1) exhibit no indication, moreover, of the formation at any intermediate stage of any conjugated di-ethenoid ester (as postulated by Boeseken *et al.* (1927, 1930) and subsequently disputed by Steger *et al.* (1944)). Since an individual conjugated diene ester of a C_{18} acid has an extinction coefficient $E_{1\text{cm}}^{1\%}$ of 1200 at 234 $m\mu$, the production of any significant proportion of such an ester would be readily seen from the graphs relating extinction coefficient to wave-length, whereas all that is apparent in figure 1 is a steady fall in the magnitude of the elaeostearate absorption bands at 268 to 290 $m\mu$ unaccompanied by any significant effect in the region 230 to 240 $m\mu$.

The significance, if any, of the minor proportions of di-ethenoid esters which appear erratically at some stages (e.g. table 2, iodine values 169.2 (110° C) and 183.3 (170° C)) is uncertain. It would not be unreasonable to attribute such instances to chance combinations of accumulated analytical error, but it is curious that direct spectrographic examination of the partly hydrogenated esters in question has also shown smaller but still significant proportions of conjugated diene unsaturation (cf. table 7). It may therefore be that at certain stages of the hydrogenation process some (relatively minor) production of di-ethenoid unsaturation from elaeostearate (or from its mixture with mono-ethenoid esters already produced from it) may take place; this point is again considered later (p. 333).

The chief primary products—mono-ethenoid esters—of hydrogenation of methyl elaeostearate are probably not homogeneous, although we can confirm the statement of Groot *et al.* that methyl octadec-11-enoate is produced in major proportions. We are of the opinion, however, that mono-ethenoid esters with the double bond in other positions (e.g. $\Delta^{9:10, 10:11, 12:13}$ or $13:14$) are almost certainly present in addition to the main product, methyl *trans*-octadec-11-enoate. We have examined in some detail the product obtained by hydrogenating methyl elaeostearate to an iodine value of about 75, when almost all elaeostearate has disappeared. The mixed acids obtained by hydrolysis of a hydrogenated ester of this type include a certain proportion of saturated acids, and it is not possible to effect a quantitative separation of the latter from the mono-ethenoid acids by crystallization. Nevertheless, systematic crystallization from acetone or ether at temperatures down to -45° C has enabled us to isolate from the mixed fatty acids about 35 to 40 % of an acid which melted at 39 to 40° C and which appeared to consist largely of *trans*-octadec-9-enoic

(vaccenic) acid. Undoubtedly this represented only a part of the total amount of this acid present, but concurrently there were definite indications of the presence of other isomeric mono-ethenoid acids.

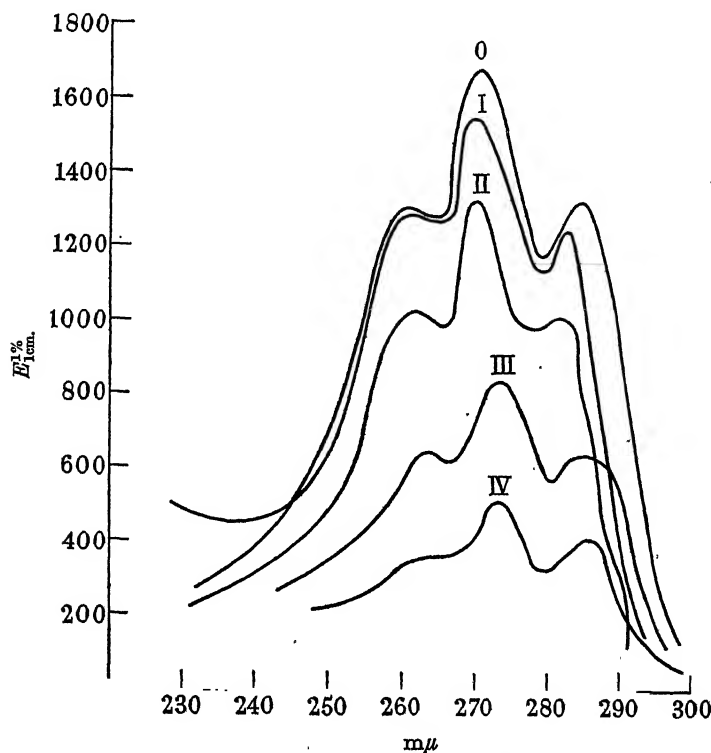


FIGURE 1. Hydrogenation of methyl elaeostearate. 0 = original ester, I to IV = hydrogenated ester. Iodine values (Toms) as follows: 0, 251.5; I, 218.0; II, 199.1; III, 169.2; and IV, 128.0.

*Hydrogenation of mixtures of methyl elaeostearate with methyl oleate,
linoleate, or linolenate*

The courses taken during the hydrogenation of equimolecular mixtures of methyl elaeostearate with either methyl oleate, methyl linoleate, or methyl linolenate were next examined under the same conditions (Raney nickel catalyst at 110 and 170° C) with the results illustrated in tables 3, 4 and 5.

Methyl elaeostearate and methyl oleate (table 3). The partly hydrogenated products were examined as in the case of methyl elaeostearate alone, i.e. the proportions of elaeostearate and of saturated esters were determined analytically, and the residual iodine value of each product calculated to a mixture of mono- and di-ethenoid esters.

Table 3 shows very clearly that no oleate is hydrogenated to stearate until all the elaeostearate has been converted into the mono-ethenoid condition, and confirms that this is the main primary process in the hydrogenation of the conjugated triene ester, although (as in the case of methyl elaeostearate alone) erratic appearances of (calculated) di-ethenoid esters occur, apparently at an iodine value of about 120. In these mixtures, however, no corresponding appearance of conjugated diene esters

was observed by direct spectrographic examination of the hydrogenated products (cf. table 8).

Methyl elaeostearate and methyl linoleate (table 4). In this series of experiments elaeostearate was determined spectrophotometrically, linoleate by similar spectrophotometric analysis after the mixtures had been isomerized with alkali at 180° C for 1 hr., and saturated components by oxidation (Bertram), the mono-ethenoid esters being determined by difference.

TABLE 3. HYDROGENATION OF METHYL ELAEOSTEARATE WITH METHYL OLEATE

iodine value (Toms)	elaestearate (spectrographic) (% found)	saturated (Bertram) (% found)	di-ethenoid (from residual iodine value) (% calc.)	mono-ethenoid (from residual iodine value) (% calc.)
hydrogenation at 110° C				
153.6	45	4	—	51
132.0	28	4	1	67
120.8	12	5	21	62
80.1	1	9	—	90
60.1	—	30	—	70
hydrogenation at 170° C				
153.6	45	4	—	51
135.3	28	4	3	65
119.3	11	5	21	63
93.7	1	6	14	79
56.7	—	32	—	68

TABLE 4. HYDROGENATION OF METHYL ELAEOSTEARATE
WITH METHYL LINOLEATE

iodine value (Toms)	elaestearate (spectrographic) (% found)	linoleate (spectrographic) (% found)	mono-ethenoid (by difference) (% calc.)	saturated (Bertram) (% found)
hydrogenation at 110° C				
193.3	45	48	7	—
159.7	25	48	22	5
131.0	8	33	54	5
92.0	1	11	81	7
61.3	—	1	71	28
hydrogenation at 170° C				
193.3	45	48	7	—
151.3	20	49	28	3
121.3	4	32	60	4
86.9	—	11	82	7
50.7	—	1	57	42

At both 110 and 170° C linoleate remained unattacked until the concentration of elaeostearate had fallen to about 20 % or less of the total esters present; thereafter, it underwent hydrogenation concurrently with elaeostearate but when the latter had practically disappeared the mixtures still contained about 25 % or more of unchanged methyl linoleate. Since the primary hydrogenation products of both elaeostearate and linoleate are mono-ethenoid esters, and hydrogenation in both

cases is extremely selective with reference to the mono-ethenoid stage, the concentration of the latter esters rises to over 80 % (or probably higher) before any substantial production of stearate sets in.

Direct spectrophotometric examination of the hydrogenated esters (cf. table 9) showed only negligible amounts of conjugated diene ester throughout the series at 110° C, but somewhat over 10 % at 170° C by the time that the linoleate was about to commence to undergo hydrogenation. This accords with the observation of Waterman & van Vlodrop (1936) that linoleate undergoes partial isomerization to conjugated forms when exposed at temperatures of about 200° C or higher to catalytically active nickel under conditions which preclude its actual combination with hydrogen.

Methyl elaeostearate and methyl linolenate (table 5). In this series elaeostearate and any small proportions of conjugated diene compounds were determined by direct measurement of the extinction coefficients at 268 and 234 $m\mu$ respectively, and linolenate was determined spectrophotometrically from the value of $E_{1\text{cm}}^{1\%}$ at 268 $m\mu$ after alkali-isomerization at 170° C for 15 min. (Hilditch *et al.* 1945). The proportions of mono- and di-ethenoid esters in the products were then estimated from the residual iodine value (cf. table 10) calculated after deducting from the total iodine value the increments due to elaeostearate and linolenate.

TABLE 5. HYDROGENATION OF METHYL ELAEOSTEARATE
WITH METHYL LINOLENATE

iodine value (Toms)	elaestearate (spectrographic) (% found)	linolenate (spectrographic) (% found)	di-ethenoid (from residual iodine value) (% calc.)	mono-ethenoid (from residual iodine value) (% calc.)
hydrogenation at 110° C				
256.7	47	48	5	—
194.8	21	33	17	29
159.9	7	20	31	42
108.1	—	3	19	78
hydrogenation at 170° C				
256.7	47	48	5	—
191.4	18	38	8	36
155.6	3	24	26	47
112.4	—	3	24	73

Mixtures of linolenate and elaeostearate esters differ strikingly from those previously discussed (elaestearate with linoleate or oleate) in that both the non-conjugated and the conjugated triene esters combine with hydrogen from the outset of the reaction. Nevertheless, the rate of hydrogenation of elaeostearate is at least twice that of the non-conjugated linolenate, the former having almost completely disappeared when about half of the linolenate has been attacked. Little if any conjugated diene ester was detected during the course of hydrogenation at either 110 or 170° C.

Behaviour of methyl elaeostearate and oleate with Raney nickel in absence of hydrogen. A possible explanation of the appearance of small proportions of diene esters during hydrogenation of methyl elaeostearate might conceivably be the donation of hydrogen by oleate or other mono-ethenoid ester to the conjugated triene ester, the diene

compounds thus arising from dehydrogenation of a mono-ethenoid ester. Methyl elaeostearate (1 part) and methyl oleate (2 parts) were therefore stirred in the hydrogenation vessel for 2 hr. with Raney nickel in an atmosphere of nitrogen. Whilst small increases in the linoleate content (determined spectrophotometrically after alkali-isomerization, table 11) were observed both at 110 and at 170° C, there was little corresponding diminution in the proportion of elaeostearate present at the conclusion of the experiments:

	elaestearate (% found)	linoleate (% found)
original	32	1
in nitrogen for 2 hr. at 110° C	31	6
in nitrogen for 2 hr. at 170° C	30	8

On the other hand, after heating with Raney nickel in an atmosphere of nitrogen at 180° C for 3 hr., no formation of diene esters was detectable (linoleate in original ester 0.5 %, after the experiment 0.4 %).

The cause of the occasional appearance of diene unsaturation in minor proportions at certain stages of the hydrogenation of elaeostearic esters has thus not been ascertained during the course of the present work.

CONCLUSIONS

We believe that the most interesting feature revealed by this investigation is the comparable reactivity to hydrogen in presence of nickel of the unsaturated systems —CH:CH.CH₂.CH:CH.CH₂.CH:CH— and —CH:CH.CH:CH.CH:CH—.

It may be deduced from the present data together with those of Bailey & Fisher (1946) for oleate, linoleate and linolenate that the relative reactivities to hydrogen of elaeostearate, linolenate, linoleate and oleate are of the order 80:40:20:1. The pentadiene carbon system in linoleic ester and especially the doubled pentadiene arrangement in the linolenic series are thus relatively much more akin to the conjugated triene grouping than to an isolated ethylenic group in their behaviour to hydrogen in presence of catalytic nickel.

The present studies have also defined more clearly the course of hydrogenation of the elaeostearic conjugated triene system, in which the primary phase has been shown to consist almost wholly in the simultaneous addition of four atoms of hydrogen to the elaeostearate molecule, with the production in one stage of a mono-ethenoid, or more probably a mixture of isomeric mono-ethenoid, esters. It has also been made clear that selectivity of hydrogenation is as well defined at the relatively low temperature of 110° C as at higher temperatures, if the nickel catalyst employed is in a form which is suitably active at a low temperature.

ANALYTICAL DATA

The analytical determinations upon which are based the results discussed in the preceding section and enumerated in tables 1 to 5 are summarized for convenience at this point (tables 6 to 11). The methods employed in the various determinations have been described earlier in this paper (cf. pp. 326 to 328).

TABLE 6. HYDROGENATION WITH METHYL LINOLEATE

at 110° C			at 170° C		
iodine value (Wijs)	linoleate (spectrophotometric, after alkali-isomerization)		iodine value (Wijs)	linoleate (spectrophotometric, after alkali-isomerization)	
	$E_{1\text{cm.}}^{1\%}$	234 m μ % linoleate		$E_{1\text{cm.}}^{1\%}$	234 m μ % linoleate
163.4	800	88.3	163.4	800	88.3
133.1	502	55.4	138.2	557	61.5
101.7	167	18.4	94.1	101	11.2
78.7	16	1.7	85.3	5	0.6

TABLE 7. HYDROGENATION OF METHYL ELAEOSTEARATE

spectrophotometric analyses (unisomerized)						
iodine value (Toms)	conj. triene		conj. diene	saturated (Bertram)	calc. iodine value of mono- + di-ethenoid	
	$E_{1\text{cm.}}^{1\%}$	268 m μ (%)				
at 110° C						
251.5	1659	93.2	155	—	—	125
218.0	1534	86.2	259	6.6	4.3	—
199.1	1294	72.7	228	6.4	7.6	50
169.2	841	47.3	212	9.4	10.1	120
128.0	477	26.8	138	6.8	13.0	101
76.3	101	5.7	162	12.0	22.2	90
50.5	1	—	—	—	50.1	106
at 170° C						
251.5	1659	93.2	155	—	—	125
219.0	1438	80.8	272	8.7	3.8	56
203.0	1194	67.1	277	11.4	4.7	104
183.3	716	40.2	270	15.5	5.9	153
113.3	279	15.7	191	13.2	9.4	101
68.7	8	—	32	2.6	29.6	94

TABLE 8. HYDROGENATION OF METHYL ELAEOSTEARATE WITH METHYL OLEATE

spectrophotometric analyses (unisomerized)						
iodine value (Toms)	conj. triene		conj. diene	saturated (%)	calc. iodine values of mono- + di-ethenoid	
	$E_{1\text{cm.}}^{1\%}$	268 m μ (%)				
at 110° C						
153.6	809	45.4	155	5.0	4.2	90
132.0	500	28.1	114	4.6	4.2	91
120.8	219	12.3	67	3.4	5.2	113
80.1	23	1.3	20	1.4	8.9	90
60.1	5	0.3	8	0.6	29.5	89
at 170° C						
153.6	809	45.4	155	5.0	4.2	90
135.3	506	28.4	146	7.2	3.9	95
119.3	201	11.3	88	5.4	4.8	112
93.7	12	0.7	20	1.5	6.4	104
56.7	3	—	7	0.6	31.9	84

TABLE 9. HYDROGENATION OF METHYL ELAEOSTEARATE
WITH METHYL LINOLEATE

iodine value (Toms)	spectrophotometric analyses						saturated (Bertram) (%)
	unisomerized			after alkali-isomerization			
	$E_{1\text{cm.}}^{1\%}$ 268 m μ	conj. triene (%)	$E_{1\text{cm.}}^{1\%}$ 234 m μ	conj. diene (%)	$E_{1\text{cm.}}^{1\%}$ 234 m μ	diene (%)	
at 110° C							
193.3	806	45.3	70	—	526	48.2	—
159.7	445	25.0	102	4.1	484	48.0	4.9
131.0	136	7.6	45	2.4	309	32.5	5.4
92.0	11	0.6	17	1.3	101	11.0	7.4
61.3	2	—	8	0.6	5	0.6	28.2
at 170° C							
193.3	806	45.3	70	—	526	48.2	—
151.3	350	19.7	180	11.6	484	49.2	2.9
121.3	83	4.6	90	6.7	296	31.7	3.8
86.9	3	—	12	1.0	104	11.4	6.9
50.7	2	—	9	0.8	11	1.2	42.3

TABLE 10. HYDROGENATION OF METHYL ELAEOSTEARATE
WITH METHYL LINOLENATE

iodine value (Toms)	spectrophotometric analyses								calc. iodine value of mono- + di-ethenoid
	unisomerized				after alkali-isomerization				
	$E_{1\text{cm.}}^{1\%}$ 268 $m\mu$	conj. triene (%)	$E_{1\text{cm.}}^{1\%}$ 234 $m\mu$	conj. diene (%)	$E_{1\text{cm.}}^{1\%}$ 268 $m\mu$	lino- lenate (%)	$E_{1\text{cm.}}^{1\%}$ 234 $m\mu$	diene (%)	
at 110° C									
256.7	839	47.1	80	—	1052	48.2	386	2.1	177
194.8	374	21.0	93	4.1	529	32.8	321	10.3	124
159.9	123	6.9	80	5.5	222	19.8	221	10.4	129
108.1	2	—	8	0.6	19	3.2	121	11.3	108
at 170° C									
256.7	839	47.1	80	—	1052	48.2	386	2.1	177
191.4	328	18.4	80	3.5	513	37.8	339	9.7	107
155.6	59	3.3	108	8.4	182	23.7	242	11.1	122
112.4	2	—	5	0.4	18	3.1	105	9.6	113

TABLE 11. METHYL ELAEOSTEARATE (1 PART) AND METHYL OLEATE (2 PARTS)
WITH RANEY NICKEL IN NITROGEN ATMOSPHERE

	spectrophotometric analyses					
	unisomerized			after alkali-isomerization		
	$E_{1\text{cm.}}^{1\%}$ 268 $m\mu$	conj. triene (%)	$E_{1\text{cm.}}^{1\%}$ 234 $m\mu$	conj. diene (%)	$E_{1\text{cm.}}^{1\%}$ 234 $m\mu$	diene (%)
elaeostearate-oleate mixture:						
original	565	31.7	60	—	76	1.5
after 2 hr. at 110° C	555	31.2	86	2.7	118	6.3
after 2 hr. at 170° C	538	30.2	151	10.9	131	7.8
methyl oleate:						
original	—	—	—	—	5	0.5
after 3 hr. at 180° C	—	—	—	—	4	0.4

REFERENCES

- Armstrong, E. F. & Hilditch, T. P. 1925 *Proc. Roy. Soc. A*, 108, 121.
Bailey, A. E. & Fisher, G. S. 1946 *Oil & Soap*, 23, 14.
Bertram, S. H. 1925 *Z. dtsch. Öl- u. Fettindustr.* 45, 733.
Boeseken, J. & Hoogland, J. 1927 *Rec. trav. Chim. Pays-Bas*, 46, 629.
Boeseken, J., van Krimpen, J. & Blanken, P. L. 1930 *Rec. trav. Chim. Pays-Bas*, 49, 247.
Groot, E. H., Kentie, A. & Knol, H. W. 1947 *Rec. trav. Chim. Pays-Bas*, 66, 633.
Harper, D. A. & Hilditch, T. P. 1937 *J. Soc. Chem. Ind., Lond.*, 56, 322.
Hilditch, T. P. 1946 *Nature*, 157, 586.
Hilditch, T. P., Morton, R. A. & Riley, J. P. 1945 *Analyst*, 70, 68.
Hilditch, T. P. & Terleski, J. T. 1937 *J. Soc. Chem. Ind., Lond.*, 56, 315.
Kaufman, H. P. & Mestern, H. E. 1936 *Ber. dtsch. chem. Ges.* 69 [B], 2684.
Moore, C. W. & Hilditch, T. P. 1923 *J. Soc. Chem. Ind., Lond.*, 42, 15T.
Moore, H. K., Richter, C. A. & van Arsdell 1917 *J. Industr. Engng Chem.* 9, 451.
Pavlic, A. A. & Adkins, H. 1946 *J. Amer. Chem. Soc.* 68, 1471.
Richardson, A. S., Knuth, C. A. & Milligan, C. H. 1925 *Industr. Engng Chem.* 17, 80.
Steger, A., van Loon, J. & van Vlimmeren, P. J. 1944 *Fette u. Seifen*, 51, 49.
Toms, H. 1928 *Analyst*, 53, 69.
Waterman, H. I. & van Vlodrop, C. 1936 *J. Soc. Chem. Ind., Lond.*, 55, 320T.
Wijs, J. J. A. 1898 *Z. angew. Chem.* 12, 291; *Z. anal. Chem.* 37, 277.

Initiation of solid explosives by impact and friction: the influence of grit

BY F. P. BOWDEN, F.R.S. AND O. A. GURTON

*Laboratory for the Physics and Chemistry of Rubbing Solids,
Department of Physical Chemistry, Cambridge*

(Received 17 January 1949)

This paper describes an experimental study of the initiation of solid explosives, and in particular the effect of artificially introducing transient hot spots of known maximum temperature. This was done by adding small foreign particles (or grit) of known melting-point. The minimum transient hot-spot temperature for the initiation of a number of secondary and primary explosives has been determined in this way. It is shown that the *melting-point of the grit is the determining factor*, and all the grits which sensitize these explosives to initiation either by friction or impact have melting-points above a threshold value which lies between 400 and 550° C. Grit particles of lower melting-point do not sensitize the explosives.

The same explosives initiated by the adiabatic compression of air required, for initiation, minimum transient temperatures of the same order as the threshold melting-point values.

The results provide strong evidence that the initiation of solids as well as of liquids by friction and impact is thermal in origin and is due to the formation of localized hot spots. There is evidence that in the case of the majority of secondary explosives which melt at comparatively low temperatures, intergranular friction is not able to cause explosion and the hot spots must be formed in some other way. With the primary explosives which explode at temperatures below their melting-points, hot spots formed by intergranular friction can be important.

INTRODUCTION

An experimental study of the initiation of explosion in liquids by impact and by friction has recently been described (Bowden, Stone & Tudor 1947). It has been shown (Bowden, Mulcahy, Vines & Yoffe 1947) that in general the initiation is not

a tribochemical one. That is to say, initial chemical decomposition is not produced by a direct rupture of the molecule, nor is the chemical excitation due to the rapid shearing of adjacent molecular layers. The mechanical energy of the blow (or of the rubbing) must first be degraded into heat to form a hot spot of small but finite size. Thermal initiation then occurs at this hot spot. It was found that with liquid, gelatinous or plastic explosives these hot spots were formed most readily by the adiabatic compression of minute included gas bubbles (Bowden, Mulcahy, Vines & Yoffe 1947), but they may also be formed by the friction between rubbing surfaces, and in extreme cases by the viscous heating of rapidly flowing explosive heating itself. If small gas bubbles are present the explosion is rendered extremely sensitive to impact, and ignition may occur with the gentlest of blows. There is some indication that initiation may occur when the pressure ratio for the adiabatic compression of a small air bubble is *c.* 30:1 (Eirich & Tabor 1948). This would give a temperature for the gas in the bubble which is about 500° C.

If the hot spot is formed by friction on a solid surface, there is evidence that the maximum temperature rise is usually limited by the melting-point of the surface (Bowden & Ridler 1936). By using surfaces of known melting-point, it is therefore possible to fix and to limit the transient temperature which can be readily generated by friction or impact. It has been found that for a liquid explosive, such as nitroglycerine, being rubbed between metal surfaces, initiation does not occur unless the melting-point of the metals is *c.* 480° C or higher. Direct measurements of the surface temperature between different metals rubbing together have also been made by using them as a thermocouple. If nitroglycerine is present between the surfaces it is again found that initiation does not occur until the local temperature at the points of rubbing contact is greater than 450° C.

Measurements by a variety of experimental methods show that the hot spots generated by friction and impact are transient and last for a short time, which depends on the experimental conditions but which is usually of the order of 10^{-3} to 10^{-4} sec. For this reason, and also probably because of their small size (recent measurements by Thomas (1949) indicate that the frictional hot spots on rubbing solids may be 10^{-3} to 10^{-4} cm. in diameter) the hot-spot temperature ($> 450^{\circ}$ C) necessary to ignite the explosive is appreciably higher than the conventional ignition temperatures, which for nitroglycerine, heated in bulk, is about 200 to 250° C.

These hot spots also play an important part in the initiation of decomposition in solid explosives. This paper will describe some simple experiments on the initiation of solid explosives by friction and in particular the effect of artificially introducing into the explosive small solid foreign particles of known melting-point. It is well known of course that the introduction of solid particles ('grit') has a sensitizing effect on explosives. The classical experiments on this are those of Taylor & Weale (1938), who studied the effect of glass and of carborundum particles on the impact sensitivity of a number of explosives. They concluded that the initiation mechanism was a tribochemical one, and it is usually considered that it is the hardness of the particle which is its most important property in determining the sensitizing effect. Recently, Copp, Napier, Nash, Powell, Skelly, Ubbelohde &

Woodward (1948) have made some measurements on the friction sensitivity in the presence of grit. They did not observe any effect unless the hardness of the particle was greater than 4 on the Mohs scale.

✓ However, if the sensitization is really due to the formation of a hot spot on the surface of the grit particle, we should expect the melting-point of the grit particle itself to be of primary importance (Bowden & Gurton 1948). The work described in this paper shows that this is indeed so for solid as well as for liquid explosives, and that unless the melting-point of the grit particle exceeds 400 to 500° C it does not, under the experimental conditions used, initiate explosion. ✓

A second paper describes a high-speed camera study of explosions of liquids and solids initiated at a hot spot. It shows how these explosions develop from a comparatively slow burning into a high-speed detonation. A third paper, by Dr Yoffe, deals more fully with the initiation of liquid explosives by the adiabatic compression of gas. It also discusses the formation of hot spots by the rapid compression of air and other gases entrapped between the crystals of a solid and the part this may play in the initiation of solid explosives.

1. INITIATION OF EXPLOSION BY FRICTION

Experimental

An apparatus was constructed in which a thin layer of solid explosive could be subjected to rapid shear while held under a considerable load. The apparatus is shown in figure 1.

A thin layer of explosive was placed on the upper surface of the bar *A*, under the steel cylinder *R*. By screwing *S*, a suitable load was applied to the explosive, and was measured by the bending of the beam *B*, or by using a loaded lever arm in place of *B*. The pendulum bob *P* was then raised to a suitable height and allowed to fall, turning about the ball-race pivot. The impact on *A* forced the bar along, since it was free to slide on the lubricated surface of the fixed block *C*. In this way the explosive was suddenly subjected to a rapid shearing action. Under these conditions the explosive was highly compressed before the experiment, and the possibility of including gas was small.

Results

When P.E.T.N. (penta-erythritol tetranitrate) was subjected to this experiment no explosions were obtained when the highest loads and rates of shear were employed. However, if a few small particles of glass were introduced into the explosive before the test was carried out, explosion was obtained in every experiment. The initial size of the particles used was of the order of 100 μ . This experiment was repeated with other grits of known melting-point. The materials used as grits were crushed minerals and crushed inorganic salts. Wherever possible the salts were fused before crushing in order to remove occluded moisture. The minerals chosen were among the softer ones (hardness on the Mohs scale 2 to 3.5), so that the difference in hardness between most of the materials used was not very great.

Table 1 gives the explosion efficiencies obtained with a number of grits together with the hardnesses and melting-points of the materials. For convenience the impact results, which will be discussed later in the paper, are included in the last two columns of tables 1 to 7. If the friction explosion efficiencies set out in column 4 are compared with the hardness values in column 2, no obvious correlation can be observed, but if they are compared with the melting-points of the grits as set out in column 3, a remarkably sharp division is apparent. All grits of melting-

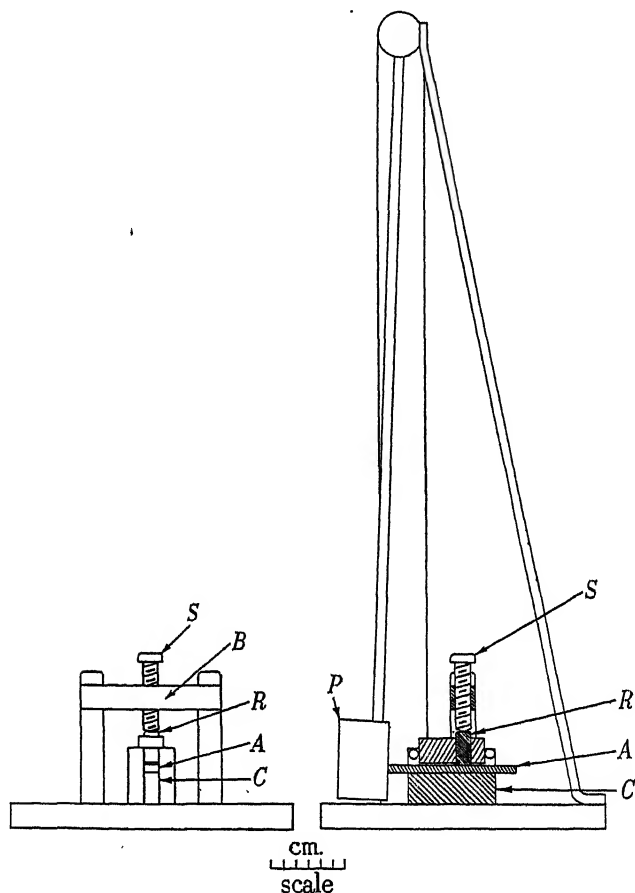


FIGURE 1. Apparatus for the initiation of explosion by friction.

point greater than 430°C were effective in causing explosion, while all grits less than about 400°C were ineffective. Since the highest temperature reached by rubbing two solids together is limited by the melting-point of the lower melting solid, it follows that the hottest spot on a piece of grit rubbed on steel, or on another piece of grit, will not, in general, have a temperature above the melting-point of the grit. Thus these experiments show that when hot spots of 430°C and upwards are produced in P.E.T.N., explosion usually follows. In the absence of any hot spot greater than 400°C no explosion occurred at the highest rates of shear used in these experiments.

Cyclonite gave results which were essentially similar to those obtained with P.E.T.N. Again grits of high melting-point were much more effective than grits of low melting-point, and no explosions were obtained if the explosive was pure, or mixed with any grit of melting-point less than 400° C. A few results have been set out in table 2.

TABLE 1. INITIATION OF P.E.T.N. BY FRICTION AND BY IMPACT IN THE PRESENCE OF GRIT

grit added	hardness	m.p. (°C)	friction	initiation	impact	initiation
			explosion no.	efficiency (%)	explosion no.	efficiency (%)
nil	—	—	0/10	0	2/97	2
ammonium nitrate	2 to 3	169.6	0/5	0	1/39	2.5
potassium bisulphate	3	210	0/5	0	1/41	2.5
silver nitrate	2 to 3	212	0/5	0	1/49	2
sodium dichromate	2 to 3	320	0/5	0	0/77	0
sodium acetate	1 to 5	324	0/5	0	0/20	0
potassium nitrate	2 to 3	334	0/5	0	0/52	0
potassium dichromate	2 to 3	398	0/5	0	0/50	0
silver bromide	2 to 3	434	5/10	50	4/65	6
lead chloride	2 to 3	501	6/10	60	8/29	27
silver iodide	2 to 3	550	5/5	100	—	—
borax	3 to 4	560	5/5	100	6/20	30
bismuthinite	2 to 2.5	685	5/5	100	5/14	42
glass	7	800	5/5	100	6/6	100
rock salt	2 to 2.5	804	5/10	50	3/50	6
chalcocite	3 to 3.5	1100	5/5	100	6/12	50
galena	2.5 to 2.7	1114	5/5	100	6/10	60
calcite	3	1339	5/5	100	6/14	43

TABLE 2. INITIATION OF CYCLONITE BY FRICTION

grit	hardness	m.p. (°C)	friction initiation	
			no.	(%)
nil	—	—	0/5	0
sodium dichromate	2 to 3	320	0/5	0
potassium nitrate	2 to 3	334	0/5	0
potassium dichromate	2 to 3	398	0/5	0
silver bromide	2 to 3	434	4/10	40
lead chloride	2 to 3	501	2/10	20
silver iodide	2 to 3	550	5/5	100
borax	3 to 4	560	5/5	100
glass	7	800	6/6	100

Experiments have been carried out on this apparatus with four initiating explosives, namely, lead azide, lead trinitroresorcinate, tetrazene and mercury fulminate, and there was one marked difference in the results. When these explosives were subjected to the stringent conditions of test used in the experiments with P.E.T.N. and cyclonite, they exploded in the absence of any grit. In consequence the load applied was reduced to 64 kg., and the maximum height of fall of the pendulum to give no explosions was determined. Table 3 shows the relative sensitivities of the four explosives.

A fundamental difference between these explosives and the secondary explosives (P.E.T.N. and cyclonite) is shown in their behaviour when heated slowly. (The secondary explosives melt before decomposing, while the primary explosives

TABLE 3. INITIATION OF EXPLOSIVES BY FRICTION IN THE ABSENCE OF GRIT

explosive	m.p. (°C)	friction experiment		explosion efficiency	
		load (kg.)	height of fall (cm.)	no.	(%)
P.E.T.N.	141	1600	70	0/10	0
cyclonite	200	1600	70	0/5	0
lead azide	> 335	1600	70	3/3	100
		64	70	1/10	10
		64	60	0/12	0
lead styphnate	> 250	64	60	8/10	80
			45	3/5	60
			40	0/15	0
mercury fulminate	> 145	64	5	1/10	10
			2½	0/20	0

TABLE 4. INITIATION OF LEAD AZIDE BY FRICTION

grit	hardness	m.p. (°C)	friction initiation (64 kg., 60 cm.)	
			no.	(%)
nil	—	—	0/12	0
silver nitrate	2 to 3	212	0/5	0
silver bromide	2 to 3	434	0/10	0
lead chloride	2 to 3	501	3/10	30
silver iodide	2 to 3	550	6/6	100
borax	3 to 4	560	5/5	100
bismuthinite	2 to 2.5	685	5/5	100
chalcocite	3 to 3.5	1100	5/5	100
galena	2.5 to 2.7	1114	5/5	100
calcite	3	1339	5/5	100

TABLE 5. INITIATION OF LEAD STYPHNATE

grit	hardness	m.p. (°C)	friction initiation (64 kg., 40 cm.)		impact initiation (240 g., 40 cm.)	
			no.	(%)	no.	(%)
nil	—	—	0/15	0	2/59	3.5
silver nitrate	2 to 3	212	0/17	0	0/20	0
potassium nitrate	2 to 3	334	0/30	0	—	—
potassium dichromate	2 to 3	398	0/20	0	—	—
silver bromide	2 to 3	434	1/35	3	0/20	0
lead chloride	2 to 3	501	4/19	21	1/55	2
silver iodide	2 to 3	550	5/6	83	5/5	100
borax	3 to 4	560	5/7	72	5/5	100
bismuthinite	2 to 2.5	685	5/5	100	5/12	40
chalcocite	3 to 3.5	1100	5/5	100	10/10	100
galena	2.5 to 2.7	1114	5/5	100	8/10	80
calcite	3	1339	14/15	93	9/11	80

investigated cannot be melted. They decompose explosively *while they are still in the solid state*. It is not unlikely, therefore, that hot spots which develop on the surfaces of these explosive crystals, or between the crystals and the steel surfaces, could reach temperatures above the minimum necessary for explosion. With the secondary explosives this will not occur, since the temperature rise is limited by the melting of the explosives.

The effect of added grits on two of these explosives was studied, and again it was found that only grits of melting-point 500°C and greater were effective sensitizers for lead azide and lead styphnate. Some results are set out in tables 4 and 5. Mercury fulminate and tetrazene were not treated in this way, since they were considered to be too sensitive to friction even in the absence of grit.

2. INITIATION OF EXPLOSION BY IMPACT

The effect of added grits on the impact sensitivity of various explosives has also been studied.

Experimental

In the first series of experiments P.E.T.N. was used. About 25 mg. of this explosive was spread as a continuous layer of crystals on the surface of a $\frac{1}{2}$ in. diameter Hoffman steel roller, and a similar roller was placed on top of the explosive. The two rollers were inserted in a $\frac{1}{2}$ in. diameter hole in a brass block mounted on a rigid table. A 225 g. Hoffman steel ball was suspended vertically above the top roller, and then allowed to fall freely. By repeating this experiment many times at various heights, a height (47 cm.) was found at which only very occasional explosions occurred. Further experiments were then carried out under these conditions, but before placing the top roller on the layer of explosive a little grit (1 or 2 mg.) was sprinkled over the explosive.

Results

The results for P.E.T.N. are recorded in table 1, columns 5 and 6. It is apparent that they are very similar to those obtained in the friction experiments. Again there is sharp division; grits of melting-point 430°C and higher are effective sensitizers, while grits of lower melting-point are in general unable to cause any increase in sensitivity.

Similar experiments were carried out with the primary explosives lead azide, lead styphnate, mercury fulminate and tetrazene. For convenience, similar rollers were used in the confining surfaces, and the hammer used was attached to an arm swinging freely about a point. Again the effective grits were those whose melting-points exceeded a certain value.

The results for lead styphnate, mercury fulminate and tetrazene are set out in tables 5, 6 and 7. It will be seen that there is again a clear-cut relation between the melting-point of the particles and its influence on sensitivity. Experiments were also carried out with lead azide. Again it was found that the melting-point was important. Silver nitrate (m.p. 212°C), silver bromide (m.p. 434°C) and lead chloride (m.p. 501°C) had no appreciable sensitizing effect when the impact was provided by a 240 g. striker falling 29 cm. Under similar conditions borax (m.p.

560° C) and chalcocite (m.p. 1100° C) gave 100 % explosion efficiency. Bismuthinite and galena had only small sensitizing effects in spite of their high melting-points, but both of these substances were soft compared with the azide.

TABLE 6. INITIATION OF MERCURY FULMINATE

grit	hardness	m.p. (°C)	impact initiation (240 g., 35 cm.)	
			no.	(%)
nil	—	—	0/20	0
silver nitrate	2 to 3	212	0/20	0
silver bromide	2 to 3	434	0/20	0
lead chloride	2 to 3	501	0/20	0
silver iodide	2 to 3	550	7/10	70
borax	3 to 4	560	5/5	100
bismuthinite	2 to 2.5	685	5/5	100
chalcocite	3 to 3.5	1100	5/5	100
galena	2.5 to 2.7	1114	5/5	100
calcite	3	1339	5/5	100

TABLE 7. INITIATION OF TETRAZENE BY IMPACT

grit added	hardness (Mohs' scale)	m.p. (°C)	impact initiation (240 g., 10 cm.)	
			no.	(%)
nil	—	—	0/101	0
silver nitrate	2 to 3	212	0/30	0
potassium nitrate	2 to 3	334	1/40	2.5
potassium dichromate	2 to 3	398	0/30	0
silver bromide	2 to 3	434	5/16	31
lead chloride	2 to 3	501	4/14	30
silver iodide	2 to 3	550	4/5	80
borax	3 to 4	560	4/4	100
bismuthinite	2 to 2.5	685	4/4	100
chalcocite	3 to 3.5	1100	3/8	38
galena	2.5 to 2.7	1114	4/4	100
calcite	3	1339	3/8	38

The chemical nature of the grit particle

An examination of the results set out in the various tables shows that the chemical nature of the grit particle has surprisingly little effect compared with its melting-point. The chemical properties of the substances used differ very widely, but in each case it is the melting-point of the particle which is the decisive factor. The one exception to this which has so far been observed is potassium chlorate. The hardness of potassium chlorate is 2 to 3 on the Mohs' scale and its melting-point is c. 370° C. This is below the critical hot-spot temperature of 400 to 500° C. Experiments showed, however, that the addition of potassium chlorate could have some sensitizing effect on the initiation of certain explosives. When used with P.E.T.N., for example, no sensitization was observed for the friction experiments, but the explosion efficiency under impact was 18 %. This should be compared with the results given in table 1. When mixed with lead azide and investigated under the

conditions obtaining in table 4 it was found that again it had no apparent sensitizing effect on the friction initiation, but the explosion efficiency under impact was 6 %. With lead styphnate (compare table 5) the explosion efficiency under friction was 7 % and with mercury fulminate under impact (compare table 6) it was 11 %. Particles of potassium chlorate (m.p. 370° C) can therefore have some sensitizing effect when substances such as potassium dichromate (m.p. 398° C) have none. When we consider the strong oxidizing properties of potassium chlorate and the fact that when mixed even with stable substances like carbon, it will readily explode on heating, this observation is not surprising. It is probable that extended and more detailed observations would yield further information on the influence of the chemical properties of the grit on sensitizing (or, for that matter, on desensitizing) the explosive. It is clear, however, from tables 1 to 7 that the overriding factor is the melting-point.

Influence of size, thermal conductivity and hardness of the grit

We should expect that the size, the hardness, and thermal conductivity of the grit particle would all play an important part because of the influence they have on the ease of the formation of hot spots. If the particles are too small and too numerous the energy of impact will be dissipated over many points of contact which are distributed over a large area, so that no single hot spot reaches the required temperature for ignition. Experiments on the frictional initiation of nitroglycerine rubbed between metal surfaces in the presence of carborundum have, in fact, shown that larger grit particles (100 μ in size) were more effective in producing hot spots than smaller particles (0.6 to 10 μ).

It is clear that although the maximum hot-spot temperature is fixed by the melting-point of the particle, the ease with which the hot spot is formed will be very dependent upon the hardness. With a hard sharp particle the stresses will be concentrated at one or two points so that it will require a much smaller energy, under conditions both of impact and friction, to produce a localized temperature rise of the necessary magnitude. If the particle is soft it will be plastically deformed or crushed so that this local concentration of the energy is not possible. For this reason we should expect that hard particles would be much more effective than soft ones provided the melting-point of the particles is above the critical value.

Equally, we should expect that the thermal conductivity of the particles would be of some importance. The general relation between the thermal conductivity and the formation of hot spots and the incidence of explosion has already been established in the earlier work (Bowden, Stone & Tudor 1947). It is much more difficult to get visible hot spots on rubbing surfaces and also to initiate the explosion of nitroglycerine if the surfaces are good thermal conductors. An interesting indication of the importance of the thermal conductivity of the grit particle was provided by rock salt. The hardness of rock salt is 2 to 3.5 and its melting-point is c. 800° C. It should therefore sensitize the explosives to impact and friction. Experiments showed that this was so. Experiments carried out with P.E.T.N. and with lead styphnate showed that grit particles of rock salt initiated the explosion. In general, however, the explosion efficiency was appreciably lower than

with the other grit particles of similar melting-point but lower thermal conductivity. For example, its explosion efficiency with P.E.T.N. under the conditions of table 1 was 50 % for friction initiation and 6 % for impact initiation. With other grits of similar melting-point but lower thermal conductivity, this explosion efficiency was about 100 % for friction initiation and about 30 to 60 % for impact.

Support for this view was given by simple experiments in which a small block of rock salt was rubbed against a rotating glass plate under a considerable load (c. 30 lb.). Any hot spots of 560° C or upwards would have been clearly visible. No such visible hot spots were detected, and the whole block soon became too hot to hold. Other relatively poor conducting materials such as calcite and glass gave distinct visible hot spots under the same conditions. It may be noted that the glass which is the hardest grit used in these experiments and which possesses a low thermal conductivity, gave with most of the explosives the highest explosion efficiency.

3. INITIATION OF EXPLOSION BY THE ADIABATIC COMPRESSION OF AIR

It is apparent that the hot-spot temperature necessary to initiate explosion under the conditions of these experiments varies somewhat with the different explosives investigated, but for all of them lies in the region of 400 to 550° C. This temperature is appreciably higher than the conventional ignition temperature obtained by dropping the explosive into a hot atmosphere or on to a hot plate. Some of these values are given in table 8, column 2. However, the explosions obtained in these determinations occurred after induction periods of at least several minutes, while the frictional hot spots were short-lived and were unlikely to last more than 10^{-2} sec. An attempt was made therefore to estimate the minimum ignition temperature of the explosives nitroglycerine, P.E.T.N., lead azide, lead styphnate, mercury fulminate and tetrazene when the high-temperature source was applied for a very short time. The apparatus employed is shown in figure 2.

TABLE 8. IGNITION TEMPERATURE UNDER VARIOUS CONDITIONS

explosive	ignition temp. (°C)	induction period (sec.)	min. hot-spot temp. (°C)		min. temp. for ignition by adiabatic compression of air
			for initiation by friction	for initiation by impact in presence of grit	
nitroglycerine	200 ⁽¹⁾	—	450 to 480 ⁽⁴⁾	—	450 to 480
P.E.T.N.	215 ⁽¹⁾	—	400 to 430	400 to 430	480 to 500
lead styphnate	250 ⁽²⁾	90	430 to 500	500 to 550	570 to 600
lead azide	335 ⁽²⁾	10	430 to 500	500 to 550	570 to 600
mercury fulminate	145 ⁽²⁾	400	—	500 to 550	630 to 690
tetrazene	160 ⁽³⁾	5	—	400 to 430	400 to 450

(1) Belayev & Yuzefovich (1940).

(2) Kast & Haid (1924).

(3) Rinkenbach & Burton (1931).

(4) Bowden, Stone & Tudor (1947).

A small quantity of explosive *E* was placed on the piston *P* which was mounted in the block *B* and could be moved up and down by the screw *S*. A larger piston *R*

was placed in the cavity *A* which was of exactly the same shape as the piston. A 225 g. ball was dropped from 180 cm. on to the piston *R*. The air in *A* was thereby compressed into the small volume *E* in about 2 msec. and then rapidly expanded. The total time of compression and rarefaction did not exceed 10^{-2} sec. In the adiabatic compression of an ideal gas,

$$\frac{T_2}{T_1} = \left(\frac{v_1}{v_2}\right)^{\gamma-1},$$

where v_1 and v_2 are the initial and final volumes, and T_1 and T_2 the initial and final temperatures and γ the ratio of the specific heats. In these experiments v_1 and T_1 were kept constant, while v_2 was varied by raising and lowering the piston *P* by means of the screw *S*. Consequently the maximum gas temperature could be calculated. The correction for the covolume of air was found to be insignificant, since the highest volume compression ratios used were only about 20:1.

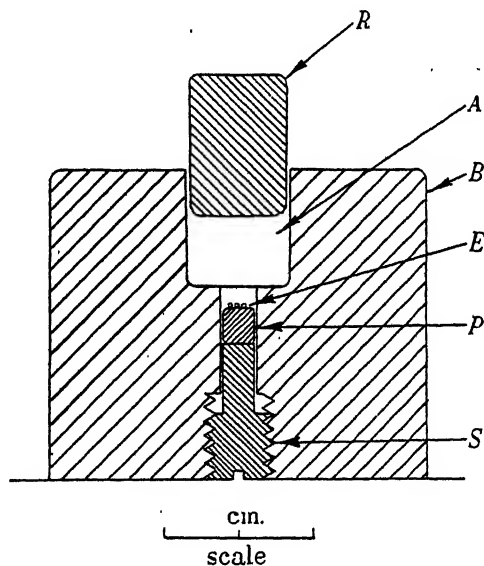


FIGURE 2. Apparatus for the initiation of explosion by the rapid compression of air.

The minimum volume ratio to give an explosion in three trials was determined as well as the maximum volume ratio to give no explosions. From these values the ignition temperature was calculated. The results have been set out in column 6, table 8, and may be compared with the values obtained for the hot-spot temperatures necessary for the initiation of explosives by friction and impact in the presence of grit which are shown in columns 4 and 5. It is interesting to note that the ignition temperatures are of a slightly higher order than the minimum hot-spot temperatures, but there is a much closer agreement between these figures than between the minimum hot-spot temperatures, and the conventional thermal ignition temperatures. Copp *et al.* (1948) calculated the order of temperature which would be required to initiate explosion by the following method.

With most explosives the thermal decomposition obeys a first-order law $k = \frac{1}{t} \ln \frac{1}{1-\alpha}$, where t is the time in seconds, α the fraction decomposed, and k the velocity constant. In order that the explosion may propagate, a finite quantity of the explosive must be decomposed within the time of application of the hot spot. Since the velocity constant must also obey an Arrhenius expression, a relationship between the time t and the temperature T can be obtained for a constant fraction decomposed. Alternatively, if t can be estimated, α can be calculated for various hot-spot temperatures.

The following values of the velocity constants have been given for the thermal decomposition of nitroglycerine, P.E.T.N., cyclonite and mercury fulminate:

nitroglycerine	$k = 10^{20.5} e^{-48,000/RT}$	(Roginsky 1932),
P.E.T.N.	$k = 10^{19.8} e^{-47,000/RT}$	(Robertson 1948),
cyclonite	$k = 10^{18.5} e^{-47,500/RT}$	(Robertson 1948),
mercury fulminate	$k = 10^{11.05} e^{-25,400/RT}$	(Vaughan & Phillips 1949).

In the next paper it will be shown that initiation of explosion may begin at 5×10^{-5} sec. from the first instant of impact. The delay between impact and the first appearance of light from the explosion includes the time of compression during which the explosive is made to flow, seal off a gas pocket, and compress the gas, as well as the time between the development of the hot spot and the onset of explosion. However, some further experiments reported in the next paper suggest that the explosion of P.E.T.N. occurs less than 10^{-5} sec. after the collapse of the gas pocket, that is, after the production of the hot spot. Using the above expressions for velocity constants, table 9 has been drawn up to show the expected proportion of each explosive decomposed in 10^{-5} sec. at various temperatures.

TABLE 9. FRACTION OF EXPLOSIVE DECOMPOSED IN 10^{-5} SEC.

temperature		nitroglycerine	P.E.T.N.	cyclonite ($\times 10^{-6}$)	mercury fulminate ($\times 10^{-4}$)
(°K)	(°C)				
600	327	0.014	0.006	2	7.2
650	377	0.240	0.120	0.005	0.004
700	427	0.980	0.620	0.056	0.015
750	477	1.000	1.000	0.430	0.051
800	527	1.000	1.000	0.980	0.14
850	577	1.000	1.000	1.000	0.31
900	627	1.000	1.000	1.000	0.57
950	677	1.000	1.000	1.000	0.95

Since these calculations involve extrapolation for outside the range of temperature over which determinations of the velocity constant have been carried out, they cannot be expected to give accurate results, but the calculated values are nevertheless in fairly good agreement with the values determined by adiabatic compression of air and by friction at solid surfaces. For 50 % decomposition the necessary temperature would appear to be about 400° C for nitroglycerine and P.E.T.N., about 480° C for cyclonite and a little over 600° C for mercury fulminate which compare with the determined values shown in table 8.

The fairly close agreement between the determined hot-spot temperatures for initiation by friction or adiabatic compression of air and the values calculated by extrapolating thermal decomposition data provide additional evidence for the suggestion that the chemical change which constitutes the onset of explosion induced mechanically is the same change as occurs at lower temperatures during slow decomposition.

It must be concluded that the mechanical initiation of these solid and liquid explosives is due to the development of local high temperatures by friction between solid surfaces, or by adiabatic compression of gas, and it is not due to direct mechanical activation of surface molecules.

We thank the Ministry of Supply (Air) for support and equipment and also Dr James Taylor (Research Manager) and Imperial Chemical Industries (Nobel Division), from whose department one of us (O.A.G.) was seconded to this laboratory.

REFERENCES

- Belayev, A. F. & Yuzefovich, N. A. 1940 *C.R. Acad. Sci. U.R.S.S.* **27**, 133.
Bowden, F. P. & Gurton, O. A. 1948 *Nature*, **162**, 654.
Bowden, F. P., Mulcahy, M. F. R., Vines, R. G. & Yoffe, A. 1947 *Proc. Roy. Soc. A*, **188**, 291.
Bowden, F. P. & Ridler, K. E. W. 1936 *Proc. Roy. Soc. A*, **154**, 640.
Bowden, F. P., Stone, M. A. & Tudor, G. K. 1947 *Proc. Roy. Soc. A*, **188**, 329.
Copp, J. L., Napier, S. E., Nash, T., Powell, W. J., Skelly, H., Ubbelohde, A. R. & Woodward, P. 1948 *Phil. Trans. A*, **241**, 280.
Eirich, F. & Tabor, D. 1948 *Proc. Camb. Phil. Soc.* **44**, 566.
Kast, H. & Haid, A. 1924 *Z. angew. Chem.* **38**, 43.
Rinkenbach, W. H. & Burton, O. E. 1931 *Army Ordn.* **12**, 120.
Robertson, A. J. B. 1948 *J. Soc. Chem. Ind., Lond.*, **113**, 221.
Roginsky, S. Z. 1932 *Phys. Z. Sowjet.* **1**, 640.
Taylor, W. & Weale, A. 1938 *Trans. Faraday Soc.* **34**, 995.
Thomas, P. H. (1949) See Bowden, F. P. & Tabor, D. *J. Inst. Mech. Eng.* (in the press).
Vaughan, J. & Phillips, L. 1949 *J. Chem. Soc.* (in press).

Birth and growth of explosion in liquids and solids initiated by impact and friction

BY F. P. BOWDEN, F.R.S. AND O. A. GURTON

*Laboratory for the Physics and Chemistry of Rubbing Solids,
Department of Physical Chemistry, Cambridge*

(Received 17 January 1949)

[Plates 11 to 14]

The birth and growth of explosions initiated by mechanical and thermal means have been studied. Liquid and solid explosives show a striking similarity. The point of initiation is always located at a source of local high temperature, for example, a hot wire, an electric spark, an impacted grit particle, or at a gas pocket suddenly compressed during impact.

There is an appreciable time lag between the first moment of impact and the first appearance of light from the explosion. With secondary explosives the time lag depends on the conditions of impact (it could be varied from 60 to 150 μ sec.), but for all the explosives studied the delays under similar conditions are approximately the same. For the primary explosives the time lags are usually much shorter, indicating that a different mechanism of initiation may be operative.

The first stage of explosion in liquids is a burning which begins slowly and *accelerates* to speeds of 500 or even 1000 m./sec. This speed may represent, in the main, a mass movement of the gas products away from the centre of explosion.

In most solid explosives (both primary and secondary) the first stage is again a slow burning which accelerates to speeds of several hundred metres per sec. A second stage of constant velocity detonation then sets in. The detonation velocity (which varies from 1100 to 2300 m./sec. according to the explosive and the physical conditions of the layer) may be identified with the low-velocity detonation in large charges, and the correct order of velocity has been explained on hydrodynamic grounds.

It is suggested that the continued propagation of the low-velocity detonation stage in a liquid is made possible by the rapid breaking up of the explosive by the detonation shock front, particularly if the liquid has a low viscosity. In more viscous liquids and solids propagation is possible only if hot-spot sources are present in the explosive. The hot spots may be developed by rapid compression of gas pockets, or, if the solid has a high enough melting-point, by intercrystalline friction.

INTRODUCTION

We have seen that the initiation of both liquid and solid explosives occurs at a hot spot of small but finite size. In this paper we will consider the growth of the explosion as it develops from the small hot spot to a high-speed detonation. An account of the development of the explosion in nitroglycerine and other liquids has been given in earlier papers (Bowden, Mulcahy, Vines & Yoffe 1947; Mulcahy & Vines 1947; Vines 1947). It was shown that after the explosion had been initiated at an electric spark or a compressed gas bubble, it developed in two stages. The first stage was a rapid burning which spread at a velocity of about 400 m./sec., but this burning did not travel more than a centimetre before a detonation set in and attained a speed of 2000 m./sec.

If the hot compressed bubble was trapped in a small hole or cavity, the general explosion was preceded by a very slow burning spreading at 20 m./sec. or less.

After about 50 μ sec. the explosion burst out of the confines of the cavity and fired the rest of the explosive.

This paper will describe first a more detailed experimental examination of the behaviour of liquids and then a study of the development of explosions in solids. It will be shown that the formation of local hot spots may play an important part in the *growth and propagation* of the explosion, as well as in its birth.

EXPERIMENTAL

With the help of Mr J. S. Courtney-Pratt the drum camera was designed and erected. In principle the camera was similar to those described by earlier workers, but was very simple in design and both easy and safe to operate. The light alloy drum 29.2 cm. in diameter was mounted on the shaft of a high-speed electric motor. The film was mounted on the inside of the drum, and a prism fixed inside the camera threw the image of the explosion on the film. With this arrangement the centrifugal forces did not tend to throw the film off the drum, but kept the film pressed hard against it. By using a short-focus lens (Zeiss Sonnar f 1.9 of 5 cm. focal length) the magnification was reduced to a minimum, and thus since the writing speed was unaffected, the slopes of the traces were increased. The camera, which was contained in a box about 15 in. cube, which could be evacuated if necessary, was bolted to the floor under a concrete table. On the table above a 5 in. square hole, there was a block of steel with a straight slit 0.125 mm. wide cut through it. The camera was focused on this slit which was so adjusted that it gave an image on the film at right angles to the direction of motion. The explosive was spread on a transparent material (armour plate glass or mica) which was put on the slit. The whole arrangement is shown diagrammatically in figure 1.

When an explosion took place the flame spread from the point of initiation, and as it progressed along the narrow area viewed through the slit, it made a record on the moving film. This film was moving at a constant speed of c. 84.5 m./sec., so that each centimetre along the film represented a time of 115 μ sec. The flame front therefore drew out a complete distance-time curve of its own movement along the slit. The slope of this trace at any point was a measure of the instantaneous velocity of propagation at the corresponding point on the slit.

Liquid explosives

In the first experiments nitroglycerine was spread as a ring on a freshly cleaved mica surface with the centre of the ring immediately above the slit. The explosive was then hit by a flat brass striker. The fall hammer used for these experiments was of the pendulum type and similar to that described by Bowden *et al.* (1947). During impact the ring of explosive flowed outwards and inwards to give a thin film of liquid containing a single central air bubble. This bubble was rapidly compressed and therefore became a hot spot which constituted the point of initiation. In this way the point of initiation was located directly above the slit, and the explosion grew radially along the slit.

A typical photograph obtained with nitroglycerine is shown in figure 2, plate 11. *A* represents the point of initiation from which the explosion spread at an increasing speed until it reached the edge of the striker at *B*. The rate of propagation of the flame in this case accelerated from about 180 to 650 m./sec., and there was no sign of any sharp discontinuity in the trace between *A* and *B*, and no appearance of detonation. Undoubtedly the flame corresponded to the 400 m./sec. stage observed by Mulcahy & Vines, but showed an *acceleration* which could not be observed at lower writing speeds. It was also obvious that this burning could propagate at least $2\frac{1}{2}$ cm. without giving rise to a detonation.

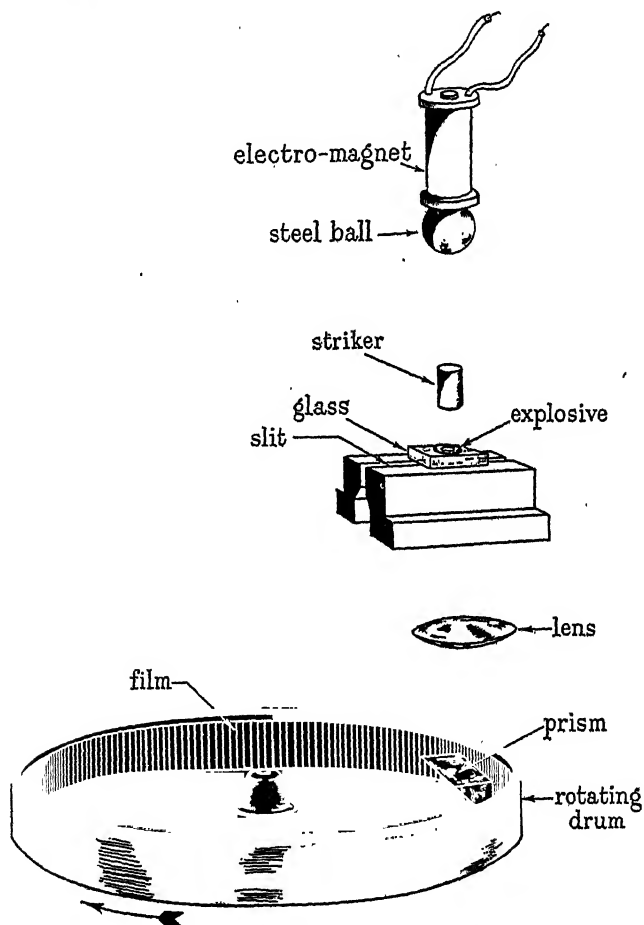


FIGURE 1. Diagrammatic representation of the high-speed camera and explosion apparatus.

When the arrangement of the explosive was changed to a ring with two projections extending along the slit, again the prolonged accelerating burning was observed, but detonation occurred before the flame reached the edge of the striker. Blast patterns on the strikers showed that the detonation usually began at points which were not directly above the slit, and this accounted for the irregular nature of the explosion photograph. Figure 3, plate 11, shows a photograph obtained this way; *AB* represented the accelerating burning. At *B* there was a dark space, some

further burning, and then a detonation. The dark space corresponded to a position at which undecomposed nitroglycerine remained on the striker. However, the dark space did not necessarily show that there was any real separation between the front of the flame and the point at which detonation set in, but merely that the transition did not occur exactly on the slit.

Initiation of a liquid explosive could also be effected by a flat striker if it had a small cavity at its centre. The cavity trapped an air bubble which was compressed by the impact, and the hot spot so produced, initiated the explosion. A thin film of nitroglycerine was spread on the transparent anvil and covered by a brass block with a hole (4.8 mm. diameter) at its centre. The explosion was then initiated by a small cavity striker in the central hole. This explosion began as a burning which continued for some time inside the cavity, and then burst out into the surrounding film of explosives. When nitroglycerine was used, and the surrounding film was confined but unimpacted except for the central 4.8 mm. the general behaviour illustrated by figure 4 was observed. In a number of experiments the rapid burning was not observed, and it is clear that detonation can occur very soon after the flame bursts out of the cavity. This is illustrated in figure 5. Similar results were obtained when cavity strikers of hemispherical section were used, but the central section of the trace was fogged by the burning of the thick layer of nitroglycerine built up round the striker during impact. A typical trace is shown in figure 6.

Initiation of explosion in nitroglycerine by electric sparks gave results similar to those obtained by Mulcahy & Vines. The spark was generated under the surface of a confined layer of nitroglycerine by bringing together a wire and thin layer of silver on the glass. The wire and the silver were previously connected to the two terminals of a charged condenser. A typical photograph in figure 7 shows that the spark initiated a rapid and accelerating burning which later developed into detonation.

Nitroglycerine could also be set off by a few crystals of lead azide initiated by a spark. In this case the explosive crystals detonated and immediately set off the nitroglycerine at its detonation velocity. This is illustrated in figure 8.

Methyl nitrate when spread as a ring and hit by a flat hammer gave photographs showing the prolonged rapid and accelerating burning from the point of initiation. Figure 9 shows a photograph obtained by the impact initiation of methyl nitrate. A novel feature of methyl nitrate photographs was the vibrating flame front of the rapid burning. This may well be connected with the high volatility of this substance which might allow alternative boiling of the liquid and combustion in the vapour phase. Methyl nitrate was also set off by cavity strikers and gave explosion traces almost identical with those obtained with nitroglycerine.

Diglycerol tetranitrate could be set off by impacting a ring of the explosive spread on mica, but the explosion which began at the compressed air bubble did not propagate far into the liquid explosive. Most of the liquid was undecomposed. Explosions initiated by cavity strikers were also very weak, and often gave insufficient light to record on the film. However, if several air bubbles were put in the liquid film, propagating explosions could be initiated by flat impact or by a cavity

striker. With cavity strikers the explosions showed the usual burning inside the cavity.

Figure 10 shows a trace obtained by initiating aerated diglycerol tetranitrate by a hemispherical cavity striker. The continued propagation was obviously a detonation, and travelled at about 2050 m./sec. When flat strikers were used the rapid burning could be detected but was of short duration and detonation set in very early.

Solid explosives

The experiments were now extended to the study of solid explosives. In all the experiments with solids the explosive layer was so arranged that the centre of the area impacted was directly above the slit, the impact was provided by a falling ball which struck a flat ended cylinder (usually $\frac{1}{2}$ in. diameter) resting lightly on the explosive. Before most experiments the layer of explosive was compressed under a load of 6000 lb./sq.in. in a hydraulic press. This was done in order to produce a layer which was uniform in thickness and as free as possible from included air. A simple device was used to measure the delay between the first instant of impact and the initiation of explosion. A spark in line with the slit was triggered by the electric contact between the falling ball and the steel cylinder. The vertical distance between the first appearance of a spark and the first light of the explosion gave a measure of the delay.

Initiation of P.E.T.N. by impact

The first explosive studied was P.E.T.N. (pentaerythritol tetranitrate). If the explosive was spread as a uniform layer of small crystals, the point of initiation was seldom located on the slit, and in many of the photographs there was evidence for several almost simultaneous points of initiation. Figure 11 shows a typical photograph obtained by impacting a thin layer of crystals 0.1 mm. thick. The long trace to the left of the explosion photograph was caused by a spark which was triggered by the impact of the falling ball on the striker (see later). However, it was found that there were at least two methods by which a single point of initiation could be obtained, and could be located directly over the slit. The first was the inclusion of a single particle of grit. A small speck of glass was placed at the centre of the continuous layer to be impacted, and then the explosion trace obtained showed the first appearance of light at the grit particle. From this point the flame spread in both directions at fairly low speeds (100 to 400 m./sec.), but showed a tendency to accelerate as it approached the edges of the impacted area. A typical trace is shown in figure 12. The fact that the explosion always began *at a grit particle* (which as we have seen is a potential source of a hot spot) provides additional evidence that the initiation is thermal.

The second method by which the point of initiation could be located was the employment of an annular distribution of the solid explosive. It has been shown that the sensitivity of a circular layer of P.E.T.N., spread on the anvil as a thin uniform layer of crystals, may be increased by removing the central portion of the layer, thus leaving a complete ring of powder (Yoffe 1948). The effect was ascribed to the consequent inclusion of a comparatively large air pocket which was sealed

off and compressed during impact. In this way a hot spot was developed by the adiabatic heating just as it was when bubbles in liquid explosives were suddenly compressed.

Further evidence for this explanation has been obtained by the use of the high-speed camera. If the ring of explosive was spread on a transparent anvil with the gas pocket directly above the slit, and struck by a hammer, the photographic trace obtained on a moving film was similar to that obtained when the explosive layer contained a grit particle. Figure 13 shows a trace obtained by impacting an annular layer of P.E.T.N. spread on mica. The explosive film was 0.1 mm. thick, and was compressed before being put in position for impact. The compression of the powder removed all large air spaces except the one purposely introduced. The energy and velocity of impact were the same as those used in obtaining the photograph shown in figure 11, and the delay between impact and initiation (as measured by the vertical distance from *S* to *A* on the time scale) was 62 μ sec., only 2 μ sec. different from the delay in figure 11. This suggests that the actual mechanism of initiation is the same for the continuous and annular layers of explosive.

Photographic observation of compression of air pocket

In order to confirm that the explosion began at the air pocket and not at some other point in the explosive layer, a new impact apparatus was constructed. This is shown in section in figure 14. The striker *H* was a disk of glass fitting into a hard steel holder. Above the striker there was a metallic mirror *M*, and by means of a lens system *L*, a beam of light from a high intrinsic brilliancy lamp *G* was directed through the glass striker and transparent anvil. The explosive was spread as a ring and placed under the striker. By an electrical mechanism the shutter was opened

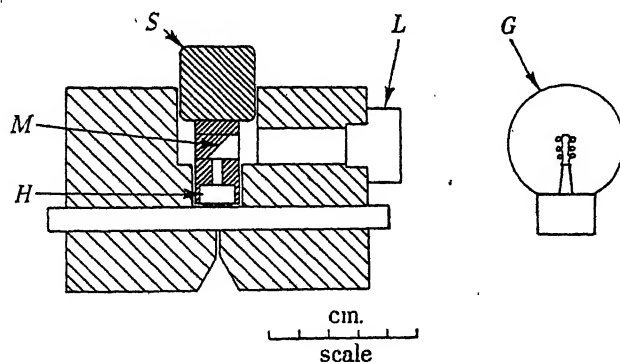


FIGURE 14. Explosion apparatus for observing the compression of the air pocket before explosion.

just before impact and closed again before a complete revolution of the drum had occurred. Thus a vertical line which was obtained on the film showed the exact position of the hole in the explosive layer (figure 15, plate 12). Less than 10^{-5} sec. before the explosion this line faded out, and the explosion always began from the position of the fade-out. The time of the fade-out may not correspond exactly with the collapse of the gas pocket, for the glass striker was always broken enough to

prevent the transmission of sufficient light, and blank experiments showed that this breaking occurred soon after the instant of impact. The experiment shows clearly, however, that initiation occurs at the gas pocket.

Delay between impact and initiation

It has already been shown that a spark trace could be put on the film at the instant when contact was made between the falling ball and the steel striker. The distance along the film from the beginning of this trace to the appearance of explosion flame was therefore a measure of the delay between impact and the first appearance of the explosion. In table 1 some delay times determined for P.E.T.N., initiated under various conditions, have been set out. It will be seen that there is in general an appreciable delay (60 to 140 μ sec.). If the explosive was spread as a ring the delay time was reduced by increasing the height of fall of the ball and thereby increasing the velocity of impact. The delay was halved by increasing the height of fall from 60 to 155 cm., in spite of a simultaneous reduction in the striker mass from 1860 to 530 g. In other words, the delay time was halved in spite of a reduction in impact energy from 11.2×10^4 to 8.2×10^4 g.cm. Thus the faster the compression the earlier was the explosion. It is therefore considered that the majority of the time between impact and explosion is taken up crushing and compressing the explosive powder and does not represent an induction period during which the explosion is undergoing some accelerating decomposition without the appearance of light.

It is also significant that the delay with a ring-like layer is about the same as the delay with a continuous film (compare rows 2 and 3 in table 1), and as will be pointed out later, the delays observed with other explosives (cyclonite and tetryl) for the same conditions of impact were very close to the figures quoted in table 1 if the same conditions of impact were observed.

TABLE 1. DELAY BETWEEN IMPACT AND EXPLOSION

mass of ball (g.)	height of fall (cm.)	conditions	delay (μ sec.)	average delay (μ sec.)
1860	60	P.E.T.N. spread as a ring	143, 135, 92, 144, 143	131
530	155	do.	79, 60, 60, 60	65
530	155	P.E.T.N. spread as a continuous layer	59, 60, 67, 62, 58	61
250	180	P.E.T.N. spread as a continuous layer	97, 137, 78	104
		P.E.T.N. containing glass particles	139, 122, 83, 78	106
		P.E.T.N. containing carborundum	91, 71, 113	90
		P.E.T.N. containing lead chloride particles	103, 74, 134, 66, 140, 136	109

A few experiments with explosive containing grit confirmed this view. There was a large scatter in the results, but there was no sign of a significantly shorter delay

when grit was present. The fourth row of table 1 shows three results obtained with flat impact on a continuous layer of P.E.T.N., and rows 5, 6 and 7 show the delay times obtained with the same set-up when glass, carborundum and lead chloride were included. There was no tendency for the delay to be reduced by the inclusion of grit particles.

Impact of single crystals

Large single crystals of P.E.T.N. can be made by slowly cooling a hot saturated solution of the explosive in acetone. The crystals are elongated plates. It was found that these single crystals could be initiated by impact. Crystals of dimensions c. $5 \times 3 \times 1$ mm. were placed on thick mica above the slit and covered by steel rollers. The explosion was then initiated by an 1860 g. ball falling 60 cm. Explosion photographs were obtained, but they were always irregular. One of these is shown in figure 16. The velocities of propagation were obviously low, but the most important observation was the fact that the delays between impact and initiation were 400 to 500 μ sec., while under the same conditions, with thin layers of the explosive the delays were 100 to 150 μ sec. It is to be concluded that initiation of the crystals did not occur until they have been crushed and broken by the impact, i.e. not until cavities and compressible gas pockets had been introduced.

Propagation of the explosion in P.E.T.N.

It has been shown that propagation from the point of initiation achieves speeds of a few hundred m./sec., but no evidence of a second stage has been mentioned.

In a series of experiments, the size of the impacted area was increased by varying the diameter of the hardened striker used. In each case the explosive was spread as a ring on the transparent anvil, the small air pocket sitting directly above the slit. Figures 17, 18 and 19 (plate 12) represent traces obtained with strikers of diameter 4.8, 12.7 and 25.4 mm. respectively. It was obvious that the propagation velocity increased as the diameter of the striker increased. Of course, there was an increase in the minimum energy necessary for initiation as the striker diameter increased, and the higher rates of propagation may have been due to the greater degree of confinement, i.e. the higher pressure of impact. However, the propagations were of a continuous nature, and no sudden change in velocity was indicated by any of the traces.

This did not apply if the area actually impacted was surrounded by a continuous layer of the same explosive. If this unimpacted area was unconfined, the explosion stopped abruptly at the edge of the striker, but if a steel or brass block was placed on top of the unimpacted portion so that it was lightly confined the explosion continued. However, it did not continue at the burning speeds of a few hundred m./sec.; there was a sudden transition to a much higher velocity. This higher velocity was accompanied by the emission of a greater amount of light. Although the velocity of propagation of this second stage of the explosion was only about 1400 m./sec., it had all the properties of a detonation wave. A typical example of this behaviour is illustrated by figure 20, which shows the explosion in a layer of explosive spread on a ring, but continued in both directions along the camera

slit. Sometimes an apparent dark space appeared where the inception of detonation occurred before the rapid burning had reached the edge of the impacted region. However, it should again be remembered that the camera observes only the phenomena which occur in the section of the explosive above the slit, and if the transition occurred at a point slightly off the slit, the detonation wave would reach the slit before the deflagration was complete because of the difference in the speeds of the two propagations. Figures 21 and 22 illustrate the two extreme examples of this. In figure 21 detonation set in immediately and there is little sign of a dark space. On the other hand, figure 22 represents the results of an experiment in which no precautions were taken to initiate the explosion on the slit. The P.E.T.N. was distributed as a continuous layer so that initiation could occur anywhere. This photograph taken at its face value suggests that the second stage or detonation begins well ahead of the advancing front of the first stage or rapid burning. However, this is illusory and merely shows that the transition has taken place first at a point off the slit, and the fast detonation wave has reached the slit before the first stage burning. If the thickness of the explosive layer was increased, the propagation velocity in the unimpacted region was increased slightly, but the velocity showed no sign of approaching the hydrodynamic detonation velocity.

Figure 23 shows a trace obtained by initiating a layer of loose crystals 0.1 mm. thick and figure 24 a trace from a layer 0.5 mm. thick. The fivefold increase in thickness caused only a 40 % increase in the detonation velocity.

The shock wave in air produced by the detonation of a thin film of P.E.T.N. could initiate a further film of the explosive. A gap of about 1.5 cm. was left in the uncompressed layer of the explosive at some distance from the point of initiation, and the whole of the explosive as before was lightly confined by a brass block. Figure 25 shows that the shock wave in air which travelled faster than the detonation itself initiated the explosive as soon as it arrived, but not at the burning speed. It is true that the new propagation started at a subdetonation speed, but it rapidly accelerated to the steady rate without any sign of discontinuity or transformation from one type of propagation to another.

Initiation by impact and propagation of explosion in cyclonite and tetryl

High-speed camera studies similar to those described above have been carried out with two other secondary explosives, cyclonite (cyclo trimethylene trinitramine) and tetryl (trinitro phenyl methyl nitramine). If either explosive was arranged as an annular layer under a striker of circular section, impact caused initiation of the explosive at the air pocket, but the propagation rate of the explosion was often irregular, and was always considerably slower than the rates observed with P.E.T.N. Figure 26 is a photograph obtained by initiating an annular layer of cyclonite by the impact of a 530 g. ball falling 154 cm. The flame persisted for a considerable time. However, the initiation delay was 61 μ sec., which was surprisingly close to the delay times observed with P.E.T.N. under the same conditions of impact (see figure 13, which shows an initiation delay of 62 μ sec.). Figure 27 is a photograph obtained by initiating a pressed film of tetryl (with a central air pocket) by the impact of an 1860 g. ball falling 60 cm. Again the pro-

pagation rate AB was much slower, but the initiation delay, $140\ \mu\text{sec.}$, was again of the same order as that observed for P.E.T.N. (120 to $150\ \mu\text{sec.}$) under similar conditions of impact.

In attempting to propagate these explosions beyond the limits of the striker's radius into an unimpacted region, results like those shown in figures 28 and 29 were obtained. Apparently in neither case was the transformation from rapid burning to detonation possible. There was some evidence of the propagation of the burning just into the unimpacted area in the case of cyclonite.

Although detonation of the unimpacted film could not be brought about by the rapid burning initiated by impact, it does not follow that thin films of these explosives cannot detonate when confined. Another set of experiments was carried out in which some lead azide was spread in the centre on the area ($1.27\ \text{cm.}$ diameter) to be impacted, while cyclonite or tetryl was spread as a continuous loose layer $0.2\ \text{mm.}$ thick in both directions along the slit from the impacted area and covered by a steel block. Initiation of the lead azide at the centre of the impacted area was assured by the inclusion of a small chip of glass. It was then obvious that these explosives could be made to detonate in thin films. Figure 30, plate 13, shows the detonation of cyclonite at $2250\ \text{m./sec.}$ and figure 31 the detonation of tetryl at 1480 and $1420\ \text{m./sec.}$ The only apparent difference between these explosives and P.E.T.N. was their inability to transform their modes of propagation from rapid burning to detonations under these conditions.

Initiation by impact and propagation of primary explosives

High-speed camera experiments have been carried out with three primary explosives, mercury fulminate ($\text{Hg}(\text{ONC})_2$), lead styphnate (lead trinitroresorcinate, $\text{PbC}_6\text{H}_3\text{N}_3\text{O}_8\text{H}_2\text{O}$), and lead azide (PbN_6). Attempts were made to initiate these explosives when spread as continuous films under $0.48\ \text{cm.}$ diameter strikers. These very small strikers were used so that there was a reasonable possibility of getting the initiation close to the slit. When thin layers of explosive and only just sufficient energy were used the traces obtained with mercury fulminate were of the form shown in figure 32. The explosion began as a rapid burning which transformed into detonation at the edge of the striker. Figure 33 is a photograph of an explosion of lead azide initiated by impacting a continuous layer of crystals by a small striker using the same conditions of experiment as have been described above for obtaining figure 32. It is apparent that the steady detonation velocity $1930\ \text{m./sec.}$ is very rapidly attained. There is no evidence for a burning and no sharp change in propagation velocity like the changes observed with P.E.T.N. and mercury fulminate.

Of course in these experiments no special precautions were taken to locate the point of initiation exactly above the slit, and consequently many of the traces were obscure in the impacted region. Some experiments were therefore designed with a view to initiating these primary explosives at known points.

In the first series of experiments the explosions were spread on the toughened glass anvils as ring-like layers with the central space over the camera slit. It was thought that this arrangement could bring about initiation by the adiabatic heating

of the air pocket trapped and compressed during impact. Figure 34 is a typical photograph obtained with mercury fulminate initiated by impacting a ring of the explosive with a striker 1.27 cm. diameter. There is no sign of a single point of initiation in this photograph, and the whole impacted region is obscure. Lead azide photographs obtained this way also showed no single point of initiation, but usually gave curved boundaries to the detonation traces indicating that initiation had occurred at a point off the slit. When lead styphnate was spread as a ring and impacted by a 1.27 cm. diameter striker, once again the point of initiation was never at the gas pocket. Figure 35 shows the photograph in which the point of initiation was nearest to the air pocket, but even here the explosion began at *A* and not at the gas pocket *P*. Thus there was no evidence for the initiation of these primary explosives by the compressional heating of gas pockets.

The only satisfactory method of ensuring the location of the point of initiation above the camera slit was the use of grit particles. Figure 36 shows a photograph from the initiation of a layer of mercury fulminate by impact when the explosive contained a tiny chip of glass at its centre. In this case the change from burning to detonation took place spontaneously at no obvious mechanical discontinuity, for the whole area of the explosive was impacted. With lead azide photographs similar to that shown in figure 27 were obtained. No burning could be detected. The behaviour of lead styphnate is illustrated in figure 38. The point of initiation was located at a grit particle on the slit, and the propagation began as an accelerating burning, but it continued into the unimpacted area without any very marked change in velocity and at an irregular speed. It must be assumed therefore either that the deflagration continued without detonation being set up, or that the detonation velocity was very low. In a second series of experiments a layer of lead styphnate was initiated by a central impacted layer of lead azide (as described for cyclonite and tetryl), and again the styphnate did not propagate at a velocity any greater than that shown in figure 38. A characteristic photograph obtained for lead azide initiation in this way is shown in figure 39.

The delays between impact and initiation when these explosives were set off by hot spots developed on glass particles proved to be irregular. This may have been connected with the uncontrolled size of the grit particles used. Here are some values obtained when the impact was provided by an arm fall hammer of effective weight 300 g. falling 50 cm.:

	$\mu\text{sec.}$
lead azide	148, 113, 68, 64, 59
mercury fulminate	107, 83, 50
lead styphnate	161, 152, 143, 129, 77

Although the scatter in these results is very wide it can be seen that although the impact energy was only 1.5×10^4 g.cm., the delays were of the same order as those obtained with P.E.T.N. using impact energies of 4.5 to 11.2×10^4 g.cm. (see table 1).

Initiation of explosion in larger crystals

Lead azide was recrystallized from a saturated solution of ammonium acetate, but since spontaneous explosions occurred in some experiments the largest crystals

grown were about 0.5 mm. long. Mercury fulminate crystals of the same size were grown by allowing a solution of the explosive in a mixture of alcohol and 880 ammonia to lose ammonia slowly. A row of lead azide crystals was placed on a sheet of armour-plate glass along the line of the camera slit, and struck by a flat brass striker weighing 200 g. and falling 40 cm. The instant of impact was recorded by an electric spark triggered by the contact of the hammer and a piece of aluminium foil. The traces obtained showed irregular intensities along their lengths and very short delays between impact and initiation (7 to 16 $\mu\text{sec.}$). A typical trace is shown in figure 40, plate 14.

When a row of mercury fulminate crystals was spread on glass, and struck by a steel striker using the fall of a 225 g. ball through 100 cm. to provide the impact, the explosion began as a deflagration. A typical trace is shown in figure 41. The burning beginning at *A* continued for only a short time. The delay between impact and initiation was also 100 $\mu\text{sec.}$, while under the same conditions of impact, fine crystals exploded and produce light 50 to 70 $\mu\text{sec.}$ after impact. It seems that in this case the development of a rapid explosion takes place only when the explosive has been crushed and a large surface formed.

Initiation of solid explosives by hot wires and electric sparks

The results of experiments on the initiation of explosives by impact were nearly all in harmony with the view that hot spots were produced by such methods as friction at a grit particle, compression of gas or vapour, or viscous flow of material, and that these hot spots were the prime cause of explosion. It was therefore important to show that if a hot spot were produced in a layer of explosive, a propagating explosion could result. It has long been known that primary explosives could be set off by hot wires or by sparks; in fact, it was this property which defined them as primary explosives. Unconfined thick trains of mercury fulminate initiated by hot wires burn at a low speed before detonation sets in, while no initial burning is found in similar experiments with lead azide (Petry 1933).

An apparatus was designed for the initiation of thin layers of explosives by hot wires. A small piece of nichrome wire was soldered at its ends to two flat strips of brass. The brass strips were bent to fit on to a piece of armour-plate glass and the glass placed on the camera slit so that the wire lay at right angles to the slit. The explosive was spread along the glass above the slit, and covered by a sheet of laminated Bakelite. By connecting the two brass strips to the terminals of a 6 V accumulator the wire was heated to bright red heat and the explosive set off.

The results obtained were very similar to those obtained in the mechanical initiation of the same explosives. Lead azide detonated immediately, mercury fulminate burned before detonating and lead styphnate explosions did not develop a speed above 700 m./sec. With mercury fulminate and lead styphnate the burning stages were of somewhat longer duration than those observed in mechanically initiated explosions.

A parallel series of experiments was carried out in which the explosives were set off by electric sparks. The results obtained were identical with the hot wire results,

a fact which suggests that the main action of the spark is to provide a hot spot and bring about thermal initiation.

A simple apparatus (shown in figure 42) was designed for initiating by an electric spark, a thin film of explosive held under confinement. The rod W and the aluminium foil strip A were connected to the terminals of a $0.05 \mu\text{F}$ condenser charged to 1000 V. The explosive film was placed on the glass G on which the aluminium foil had been stuck, and by applying a load to W the wire and foil were brought close enough to allow a spark discharge through the explosive above the camera slit.

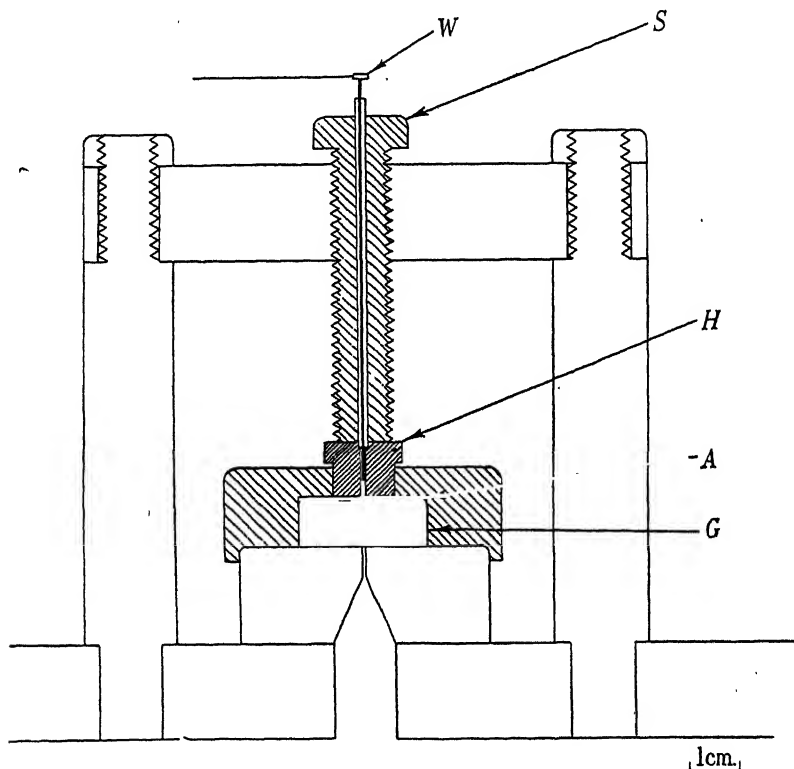


FIGURE 42. Spark initiation apparatus.

The results for lead azide initiated by a hot wire and by a spark are shown in figures 43 and 44 respectively. The photograph from spark initiation was obtained by using a very thin layer of this salt, and it revealed a very short but real delay between the passage of the spark and the onset of detonation. This delay was always observed, but it was only of the order of 1 to $4 \mu\text{sec}$.

The very different behaviour of mercury fulminate is illustrated in figure 45. Here the propagation began as a slow burning which suddenly accelerated to about 400 m./sec. When the burning reached the discontinuity between the central Bakelite insulator and the steel block surrounding it, detonation was set up. If the discontinuity was removed by using a glass insulator in the centre, which was cemented into the steel, and lapped flat, then traces of the type shown in figure 46

were obtained. The faster deflagration now continued for some distance (2 cm.) before detonation began. A typical result with the spark initiation of lead styphnate is shown in figure 47.

If the thin layers of the explosives were unconfined, the burning of mercury fulminate and lead styphnate were too slow and too weak to be recorded on the film, but lead azide was detonated immediately and the photographs could only be distinguished from those obtained with confined layers of the same explosive by the slightly lower velocity of detonation. Obviously the degree of confinement, i.e. the pressure during the deflagration stage, determined the rate of propagation of flame, and this may explain the difference between the photographs obtained by the impact initiation, and spark initiation of mercury fulminate (figures 32 and 45). During impact the high pressure set up favours an acceleration of the burning rate, and hence detonation is begun at an earlier stage.

It has long been realized that primary explosives may be initiated by electric sparks. It has already been shown that nitroglycerine when confined as a thin film may be initiated by a spark discharge but no evidence was available which indicated that solid secondary explosives could be initiated in this way. It was found that P.E.T.N. could not be initiated by heavy sparks unless the explosive layer was held under compression while the spark was passed. This was done by turning the screw *S* shown in figure 42 so that the central portion was held under compression beneath the insulator *H*, and the remainder of the explosive film was left confined but unconstrained. Sparks from a 0.05 μ F condenser charged at 100 V were then capable of initiating the explosive. Figure 47 is a photograph obtained by initiating a film of P.E.T.N. in this way; in this particular experiment a much heavier spark (1 μ F at 2000 V) was used. This photograph may be compared with the photographs obtained by the impact initiation of the same explosive, figures 20 and 21. It is very similar provided the light from the electric spark itself is ignored.

DISCUSSION

Liquid explosives

The experiments with nitroglycerine have confirmed the earlier work. The explosion of a thin film of this substance could be initiated at a rapidly compressed gas pocket, or at an electric spark discharge. In each case the explosion began as a burning, and detonation developed at a later stage. If the initiation was effected by a cavity striker the first stage of the explosion was for some time (up to 50 μ sec.) confined to the small quantity of liquid sealed off by the striker. The burning velocity was of the order of 10 to 20 m./sec. or less. When the explosion burst out of the cavity it usually produced a rapid burning in the bulk of the film and the explosion was later transformed into a detonation at about 2200 m./sec. With spark initiation or initiation by the compression of an air bubble trapped in the main body of the film during impact, only two stages of explosion could be distinguished, the initial rapid burning, and the subsequent detonation.

It has now been shown that the rapid burning stage does not show a constant propagation speed, *but the flame front accelerates as it advances*. The distance through

which the flame travels before detonation begins depends on the experimental conditions. The high initial pressure in the film during impact does not favour a transition to detonation, and the rapid burning may continue for at least 2.5 cm. In spark initiated explosions where the film is at atmospheric pressure, detonation usually occurs before the flame of the rapid burning stage has propagated more than 0.5 cm.

Methyl nitrate behaved in a manner similar to nitroglycerine although the rapid burning stage showed a vibrating front. Diglycerol tetranitrate was easily initiated by impact if gas spaces were present, but the explosion did not propagate far unless the whole film was first filled with tiny air bubbles. If this was done the onset of detonation soon occurred and the explosion was complete.

The 'dark space' (which was usually observed between the rapid burning and the detonation, and sometimes corresponded to an area on which undecomposed explosive was left after the explosion) became larger as the distance to which the burning continued increased. If the transition from burning to detonation actually occurred at a single point on the flame front the chances of this point being on the slit would be very small, and would decrease as the radius of the flame front increased. Since the detonation often travelled at speeds five times as great as the rapid burning it might be expected to reach the slit first and thus give a dark space in the high speed photographs. If the initiation is located on the slit by means of a grit particle or an air bubble, and the burning time is short, the dark space is negligibly small.

The burning speed of nitroglycerine at low pressures has been determined by Andreev (1946) and found to be a linear function of the pressure. At higher pressures Muraour (1942) has shown that the burning rate of propellants is proportional to the pressure. During impact very high pressures are developed, but they are unlikely to exceed the flow pressure of the metal. Calculating the burning speed of nitroglycerine by extrapolating Andreev's figures to 10^4 atmospheres (the approximate flow pressure of hard brass) we obtain a theoretical burning speed of about 10 m./sec. which is of the same order as the speed observed inside the brass cavity strikers. However, when the explosion flame bursts out into the main film of explosive, it attains speeds of 400 to 600 m./sec. although blast patterns on the brass confining surfaces show that the pressure is not much greater than 10^4 atmospheres. This suggests that the whole front must be undergoing a mass movement which is superimposed on the true burning speed, and attains speeds 40 to 60 times as great. In other words the products and the explosive are being forced away from the centre of the explosion.

It has in fact been shown by direct photographic measurement that liquid between the impacting surfaces may easily attain a speed of 100 m./sec. even when no explosion occurs (Bowden, Mulcahy, Vines & Yoffe 1947).

As the mass movement accelerates it may attain speeds approaching the velocity of sound in the thin film, and will therefore build up a shock front. It also attains a speed of the same order and same direction as the streaming velocity of the products behind a low velocity detonation wave. The development of low velocity detonation from the rapid movement of the flame front is therefore not surprising.

The solid secondary explosives

The results described in this paper all show how closely the behaviour of solid explosives resembles that of liquids (see also Bowden & Gurton 1948). The point of initiation in a layer of P.E.T.N. is located at a hot spot source which may be a grit particle, a compressed gas pocket, or, under suitable conditions, an electric spark.

The first stage of explosion is a rapid burning similar to that obtained in experiments with liquids. It shows an accelerating front, but no spontaneous change from burning to detonation is observed in explosions initiated by impact unless some discontinuity is present. However, if the explosive which is impacted is surrounded by an unimpacted but confined layer of the same material, detonation occurs in the unimpacted zone. The transition from burning to detonation takes place at the discontinuity, that is, at the edge of the impacted area.

Cyclonite and tetryl are similar to P.E.T.N. in their behaviour, but the flame rates in the rapid burning stages are much lower. With these explosives the transition from burning to detonation was not observed even when a discontinuity was present. Experiments showed however that if the initiation is sufficiently powerful it is possible to get a detonation which continues to propagate. If a layer of lead azide crystals is exploded in contact with a thin confined layer of cyclonite or tetryl, a propagating detonation is set up.

Once again in solid explosives the rapid burning is too fast to be accounted for by the rate of combustive consumption of the explosive under pressure. It is suggested therefore that it again represented a mass movement of the flame front, that is a movement of the gas products away from the centre of explosion. The combustion in the impacted zone builds up a high pressure, but no steep shock front is developed since the pressure of impact is itself very high. However, when the explosion bursts out from the central impacted zone the surrounding explosive received a sudden jet of high pressure gas at a very high temperature. This would be expected to produce an intense shock in the explosive, and apparently in the case of P.E.T.N. this shock brings about detonation.

The solid primary explosives

The point of initiation in the primary explosives lead azide, mercury fulminate and lead styphnate is normally located at a hot spot source, e.g. a hot wire, an electric spark or an impacted grit particle. However, an occluded air pocket does not appear to act as an initiator in impact. This is an interesting observation. With the secondary explosives studied the melting point is below the 'hot spot decomposition temperature'; they can be melted, and under impact can flow and seal off air pockets. The primary explosive cannot be melted. Their ignition temperatures are below their melting points. They are unlikely to melt and flow during impact, and the air pocket may be difficult to seal off. However, the high rates of friction and shear may lead to the development of hot spots of sufficient temperature to initiate explosion without producing melting. Thus, before the air pockets can be sealed the explosion may begin from hot spots produced by friction between crystals or between the confining surfaces and the explosive

crystals. With those secondary explosives which melt without appreciable decomposition, the temperatures of the hot spots developed by intercrystalline friction cannot exceed the melting points and therefore will not be able to initiate explosion.

The first stage of propagation, for both mercury fulminate and lead styphnate, is a rapid burning which usually shows an accelerating flame front. Pure lead azide does not show this stage of explosion, but detonation begins very near to the hot spot source. However, in spark initiation there is a short delay (1 to 4 μ sec.) between the passage of the spark and the beginning of detonation, and it is probable that this represents a very short lived burning stage.

Detonation usually begins in mercury fulminate at a mechanical discontinuity such as the junction between the impacted and unimpacted zones, but this is not always the case. It is probable that the mechanism by which a detonation wave is set up is similar in this case to the mechanism suggested for initiating detonation in liquid explosives. The rapid burning may involve the mass movement of material away from the centre of explosion which sets up a shock front in the undecomposed material.

In lead styphnate the average velocity with thin films could attain a value of about 700 m./sec. and the flame front did not show a steady speed. It is unlikely that this is a detonation.

Delay between impact and explosion

In the experiments which were designed to measure the delay between the first moment of impact and the first appearance of explosion flame, it was shown that the delay depended on the impact energy and was almost identical with P.E.T.N., cyclonite, and tetryl if the same conditions of impact were used (figures 13, 26 and 27, plates 11 and 12). For P.E.T.N. for example, under conditions of row 1, table 1, the delay was 131 μ sec., while at higher rates of impact (row 2) delays of about 65 μ sec. were obtained. With other secondary explosives such as cyclonite and tetryl the delays were very similar. With tetryl, for example, under conditions of row 1, the delay was 140 μ sec., with cyclonite it was 130. Under the conditions of row 2 the delay with cyclonite was 61 μ sec. This result is different from that obtained by Rideal & Robertson (1948) who found delay times of the order of 230 to 340 μ sec.

They measured the delay between impact and the arrival of ionized gas at a point 2 mm. outside the explosion film and found that this delay increased as the sensitivity of the explosive decreased. That is, in the order P.E.T.N., cyclonite, tetryl. However, the results of Rideal & Robertson are not incompatible with our observations, for their delay times included both the mechanical delay between impact and onset of explosion, and the time taken for the explosion to reach the edge of the impacted zone, and cause the ionizing flame to shoot out to the receiver. It has already been shown that the initial speed of propagation is much slower in cyclonite and tetryl than it is in P.E.T.N. The values obtained by Rideal & Robertson are therefore generally higher than the values obtained by high speed photography, and increase as the flame rate decreases in the explosive investigated. The con-

siderably shorter delay when grit was added to the explosive (reported by Rideal & Robertson) was not observed. The delay times which we have observed for the primary explosives, lead azide, mercury fulminate and lead styphnate were much shorter and more variable than the delays for secondary explosives. This provides further evidence for a different mechanism of initiation in these primary explosives, and again suggests that intercrystalline friction may be producing hot spots before the pressure has been raised high enough to produce any initiating hot spot by the adiabatic compression of gas.

Large crystals of P.E.T.N. were initiated only after delays of the order of 400 to 500 μ sec. which were 3 to 4 times as long as the delays observed with thin layers of small crystals. This indicates that before initiation could take place the explosive had to be crushed and compressed. The very different result with large crystals of lead azide which showed a very short delay may be explained by the hypothesis that the explosion can be initiated by intercrystalline friction and consequent hot spot formation.

Low-velocity detonation

The detonation stages observed in thin films of explosives are characterized by velocities varying from 1100 to 2500 m./sec. (neglecting the propagation of lead styphnate at 700 m./sec. which is probably a rapid burning). These values are much lower than those observed in large scale detonations (namely, 4500 to 8000 m./sec.) and correspond with the low velocity detonation characteristic of nitroglycerine explosives which has also been observed in methyl nitrate, nitroglycol and recently in solid explosives like T.N.T. and tetryl (Jones & Mitchell 1948). If the expansion of the gas product is taken into account, these low velocities can be justified on hydrodynamic-thermodynamic grounds without postulating any detailed mechanism for the propagation.

If the Abel equation of state.

$$p(V-b) = n_2 RT_2 \quad (1)$$

is applied to the reasoning of the Chapman-Jouget hydrodynamic theory of detonation the following expression for the detonation velocity is obtained (Paterson 1948)

$$D = \frac{V_1}{V_1 - b} (\gamma + 1) \sqrt{\left(\frac{n_2 RT_2}{\gamma} \right)}, \quad (2)$$

where V_1 is the volume of a gram of the explosive, b is the covolume, n_2 the number of moles of explosion products per gram, R the molar gas constant, T_2 the detonation temperature and $\gamma = 1 + n_2 R/C_2$. Here C_2 is the specific heat of the products at constant volume.

This formula would apply only if no expansion of the products took place before the reaction was complete. It can be shown by simple reasoning that if an expansion does occur the net effect on equation (2) is to increase the effective value of V_1 by a factor depending on the amount of expansion, and to decrease T_2 by a much smaller factor (never more than 10 %) so that if the expansion is large (as it probably is when thin layers of explosive detonate), V_1 becomes so large that b is

negligible in comparison. Then the term $V_1/(V_1 - b)$ tends to unity, and so equation (2) becomes

$$D = \frac{\gamma + 1}{\gamma} \sqrt{(\gamma n_2 R T_2)} \quad (3)$$

that is

$$D = \frac{\gamma + 1}{\gamma} C,$$

where C is the velocity of sound in the products. Since $(\gamma + 1)/\gamma$ is slightly less than 2 we should expect to obtain a reasonably stable minimum velocity of detonation which is approximately twice the velocity of sound in the products. In addition to this there will, of course, be the normal maximum hydrodynamic velocity. Calculations based on equation (3) have given the values for the stable minimum velocity shown in column 3 of table 2. Experimental values in column 2 were obtained by plotting a frequency curve of a large number of determinations and recording the commonest value. It is interesting to see that the results calculated by this simple theory for the velocity of detonation in thin films of explosives are in reasonable agreement with the observed values for both solids and liquids.

TABLE 2. DETONATION VELOCITIES IN THIN FILMS

explosive	velocity observed (m./sec.)	velocity calculated (m./sec.)
nitroglycerine	2200	2240
P.E.T.N.	1450	2380
cyclonite	2300	2340
tetryl	1500	2050

In P.E.T.N. and tetryl it is probable that the explosion products are different from those obtained in a high velocity detonation. The fact that the velocity in nitroglycerine (Mulcahy & Vines 1947) and in P.E.T.N. is not greatly increased by considerable changes in film thickness and in bulk density provided further support for this theory.

Although the modified hydrodynamic theory gives the correct value for the detonation velocity, it does not explain how the explosive reactions are completed in the short times available. This work provides some evidence for two possible mechanisms for this which may operate under different conditions.

In high velocity detonation the shock wave is sufficiently intense to raise the temperature of the condensed explosive to a high value by adiabatic compression (for example 3000° C with nitroglycerine (Ratner 1947)) so that the reaction can continue. In the low velocity detonation, however, the shock wave could not raise the temperature more than about 40° C. It is suggested therefore, that in the detonation of nitroglycerine and methyl nitrate, it is the mechanical action of the shock wave which facilitates propagation. That is to say the shock front breaks up the liquid into small droplets which are thrown into the reaction zone and present a large burning surface and hence a rapid reaction can ensue. Evidence for a mechanism of this type is provided by the marked effect which an increase in viscosity has on the ease of the propagation. Thus it was not found possible to

detonate a thin film of the viscous diglycerol tetranitrate, and it has been shown elsewhere that air free nitroglycerine made viscous by the addition of a very small amount of nitrocotton will not detonate at low velocity.

When the explosive is present as a thin film between solid surfaces, its break up into droplets may be facilitated by the separation of the surfaces. This is possible because the speed of sound in the solid is greater than the detonation velocity so that elastic waves in the metal due to the shock of the explosion will travel ahead of the detonation front. Some indication that standing waves might be set up in the metal which confines the explosive has been obtained in earlier work (Bowden, Eirich, Mulcahy, Vines & Yoffe 1943) where the blast marks of the explosion showed an alternating pattern on the confining metal. This break up of the explosive into droplets would also offer a reasonable explanation of the pitting and pock marks produced on brass during the detonation.

There is, however, another mechanism by which the high rate of reaction may be maintained. If small bubbles of air are present or are introduced into an explosive, although the detonation pressure would raise the temperature of the liquid or solid phase less than 40 degrees, the adiabatic compression of the gas bubbles would produce local hot spots of several thousands of degrees. These hot spots would become new explosion centres. If the air was distributed through the explosive, the detonation wave would have at its head a shock front which would therefore become a wave of initiation of new explosions. The detonation of a thin film of diglycerol tetranitrate which occurs only when air bubbles have been introduced supports this view. The well known phenomenon of 'dead pressing' in solid explosives, and the insensitiveness of old blasting gelatine and of cast explosives would all, on this view, be due simply to the removal of air. Figure 22, plate 12, shows the effect of dead pressing. Here the central region burned, but the explosion did not reach the unimpacted area above the slit before detonation had been set up. However, the detonation wave did not enter the impacted area, presumably because the pressure of impact removed most of the air and compressed the rest. In high melting primary explosives it is possible that hot spots formed by crystal fracture and intercrystalline friction in the detonation wave front may be hot enough to act as new explosion centres. Thus, there is evidence that the formation of hot spots, which has been shown to be the cause of initiation of explosion by impact and friction, *is often the means by which low velocity detonation is maintained*. Once again the commonest source of hot spots is the rapid compression of gas pockets, which are present in the explosive, but with the primary explosives hot spots may be formed by friction.

We thank the Royal Society for a grant for the development of the high speed camera, and the Ministry of Supply (Air) for support and for other equipment. Our thanks are also due to Dr James Taylor and Imperial Chemical Industries (Nobel Division) from whose Research Department one of us (O.A.G.) was seconded to this laboratory.

REFERENCES

- Andreev, K. K. 1946 *C.R. Acad. Sci. U.R.S.S.* **51**, 29, 123.
 Bowden, F. P. & Gurton, O. A. 1948 *Nature*, **161**, 348.
 Bowden, F. P., Mulcahy, M. F. R., Vines, R. G. & Yoffe, A. 1947 *Proc. Roy. Soc. A*, **188**, 291, 311.
 Bowden, F. P., Eirich, F., Mulcahy, M. F. R., Vines, R. G. & Yoffe, A. 1943 *Res. Bull. Coun. Sci. Industr. Res. Aust.* no. 173.
 Jones, E. 1928 *Proc. Roy. Soc. A*, **120**, 603.
 Jones, E. & Mitchell, D. 1948 *Nature*, **161**, 98.
 Mulcahy, M. F. R. & Vines, R. G. 1947 *Proc. Roy. Soc. A*, **191**, 210, 226.
 Muraour, H. 1942 *Chim. et Industr.* **47**, 602.
 Muraour, H. 1943 *Chim. et Industr.* **50**, 105.
 Paterson, S. 1948 *Research*, **1**, 221.
 Patry, M. 1933 Thesis, Nancy.
 Ratner, S. 1947 *Acta Physicochim. U.R.S.S.* **22**, 357.
 Rideal, E. K. & Robertson, A. J. B. 1948 *Proc. Roy. Soc. A*, **195**, 135.
 Vines, R. G. 1947 *Nature*, **160**, 400.
 Yoffe, A. 1948 *Nature*, **161**, 349.

DESCRIPTION OF PLATES 11 TO 14

PLATE 11

- FIGURE 2. High speed camera photograph of an explosion of a ring of nitroglycerine set off by impact. The point of initiation *A* corresponds to a compressed air bubble, and the explosion *AB* spreads as a rapid accelerating burning at 180 to 650 m./sec.
- FIGURE 3. Photograph similar to figure 2 for which the explosive was spread as a ring with two projections. The rapid burning *AB* gives rise to a much faster detonation at *B*.
- FIGURE 4. Explosion of a confined film of nitroglycerine by the impact of a cavity striker. Initiation occurs inside the cavity at *A*, and the slow burning *AB* (c. 10 m./sec.) is followed by a rapid burning in the confined film *BC*. Detonation *CD* at about 2200 m./sec. is the first stage in the explosion. Detonation waves *CE* travel back from the points at which detonation is first observed.
- FIGURE 5. Explosion of nitroglycerine under conditions similar to those used for figure 4. In this photograph the rapid burning stage is not apparent.
- FIGURE 6. Photograph similar to figure 4, using a hemispherical ended cavity striker. The rapid burning stage *BC* is somewhat obscured by optical effects associated with the geometry of the apparatus.
- FIGURE 7. Explosion of confined nitroglycerine initiated by an electric spark *A*, showing the rapid burning stage *AB* followed by detonation *BC*.
- FIGURE 8. Explosion of a confined film of nitroglycerine initiated by the detonation of a few crystals of lead azide at *A*. Only the detonation stage *AC* is observed.
- FIGURE 9. Explosion of a ring of methyl nitrate initiated by flat impact. Initiation occurs at a confined gas pocket *A*, and the rapid burning stage *AB* shows an accelerating, but vibrating flame front. *BC* represents the detonation stage which occurred after the burning speed had reached 1000 m./sec.
- FIGURE 10. Explosion of a confined aerated film of diglycerol tetranitrate by the impact of a cavity striker. *AB* represents the slow burning inside the cavity (53 μ sec.). Between *B* and *C* (36 μ sec.) there was no sign of explosion. The central region of the general explosion is obscure, but the detonation stage *DE* can be clearly distinguished.

FIGURE 11. Explosion of pentaerythritol tetranitrate (P.E.T.N.) by flat impact on a layer of crystals. The beginning of the spark trace *S* represents the first instant of impact. The first appearance of the explosion light occurs at *A* but there is no single point of initiation. The impact was provided by a 530 g. ball falling 155 cm.

FIGURE 12. Explosion of P.E.T.N. initiated by impact with a grit particle at *A*. The explosion began at the grit particle and spread as a rapid burning *AB*.

FIGURE 13. Explosion of a layer of P.E.T.N. spread as a ring, and initiated by a flat impact. Initiation occurred at the gas pocket *A* and spread as a rapid burning *AB*. The delay between impact and explosion was similar to that observed in figure 11, for the conditions of impact were the same. The average flame speeds were 435 and 350 m./sec.

PLATE 12

FIGURE 15. Explosion of P.E.T.N. spread as a ring and impacted by a flat glass striker. *S* is a spark trace produced at the instant of impact and about 55 μ sec. later a beam of light passing through the central gap in the explosive is extinguished at *C*. The explosion begins at *A* less than 10 μ sec. later, and spreads as a rapid burning *AB*.

FIGURE 16. Explosion of a single crystal of P.E.T.N. initiated by impact.

FIGURE 17. Explosion of a ring of P.E.T.N. by the impact of a 0.48 cm. diameter striker.

FIGURE 18. Explosion of a ring of P.E.T.N. by the impact of a 1.27 cm. diameter striker.

FIGURE 19. Explosion of a ring of P.E.T.N. by the impact of a 2.54 cm. diameter striker.

FIGURE 20. Explosion of P.E.T.N. initiated by impact on a ring of explosive propagated into a surrounding layer of confined but unimpacted explosive. In the impacted area only the rapid burning stage *AB* is observed but the surrounding explosive detonates *BC*. Retonation waves *BD* travel back into the burnt gas of the first stage of explosion.

FIGURE 21. The same as figure 20, showing distinct acceleration in the first stage, and very little 'dark space' at *B*.

FIGURE 22. Explosion of P.E.T.N. initiated by impact on a continuous layer surrounded by a layer of confined but unimpacted explosive. The first appearance of explosion light in the impacted region occurs at *A*. Between *A* and *B* (where detonation is set up) there is a distinct 'dark space'.

FIGURE 23. Detonation in a layer of P.E.T.N. 0.1 mm. thick of density 0.5 g./cm.³ at 1100 and 1200 m./sec.

FIGURE 24. Detonation in a layer of P.E.T.N. 0.5 mm. thick of density 0.75 g./cm.³ at 1575 and 1680 m./sec.

FIGURE 25. Initiation of a thin layer of P.E.T.N. by an explosion generated shock wave. The detonation *BC* gave rise to a shock wave in a layer of air *CE* which on meeting another layer of explosive set up an immediate detonation.

FIGURE 26. Explosion of cyclonite initiated by flat impact on a ring of the explosive. The explosion developed is a burning which is much slower than the burning displayed by P.E.T.N. *S* represents the instant of impact and *A* the point of initiation corresponding to a compressed gas pocket.

FIGURE 27. Explosion of tetryl initiated by flat impact on a ring of the explosive. The explosion develops as a rapid burning, but again the rate of burning is much slower than that observed in P.E.T.N. *A*, the point of initiation is located at a compressed gas pocket.

FIGURE 28. An attempt to propagate the explosion of cyclonite into a surrounding layer of the explosive. *BC* represents a short lived burning developed in the surrounding layer, but no detonation is observed.

FIGURE 29. A similar unsuccessful attempt with a layer of tetryl.

PLATE 13

FIGURE 30. Detonation of cyclonite by a layer of lead azide. *AB* represents the detonation of lead azide, and *BC* the detonation of a confined layer of cyclonite.

FIGURE 31. Detonation of tetryl by a layer of lead azide. *AB* represents the detonation of lead azide, and *BC* the detonation of a confined layer of tetryl.

FIGURE 32. Explosion of mercury fulminate initiated by impact. The explosion began as a burning *AB*, 270 to 300 m./sec. and at the edge of the impacted region transformed into a detonation *BC* at about 1600 m./sec. The light produced by the detonation of mercury fulminate was always strongly actinic and in consequence the traces were never very sharp. In this particular photograph the detonation *BC* shows some deceleration. This behaviour was exceptional and was due to the fact that in this experiment the explosive layer was so thin that all the crystals were not in contact.

FIGURE 33. Explosion of lead azide initiated by impact. No initial rapid burning can be distinguished.

FIGURE 34. Explosion of lead styphnate by impact on a ring-like layer. The point of initiation *A* is not located at the gas pocket. The explosion propagates as a rapid burning *AB*.

FIGURE 35. Explosion of a ring of mercury fulminate initiated by impact. No single point of initiation can be distinguished.

FIGURE 36. Impact initiation of a continuous layer of mercury fulminate containing a grit particle at *A*. From *A* the explosion spreads as a rapid burning *AB* which gave rise to a detonation *BC* (1450 m./sec.).

FIGURE 37. Initiation of lead azide by impact in the presence of a grit particle. Initiation occurs at the grit particle *A*, but only the detonation stage *ABC* is observed. The speed was about 2300 m./sec.

FIGURE 38. Explosion of lead styphnate initiated by impact in the presence of a grit particle *A*. From the point of initiation *A*, the explosion spreads as a rapid burning *AB* which continues in the unimpacted region at roughly the same speed *BC*.

FIGURE 39. Initiation of lead styphnate by lead azide. The detonation of lead azide *AB* gives rise to an explosion *BC* in the lead styphnate at 700 m./sec. which is much slower than the expected detonation velocity.

PLATE 14

FIGURE 40. Explosion of large crystals of lead azide initiated by flat impact. Note the very short delay ($< 10 \mu\text{sec.}$) between impact *S* and explosion *A*.

FIGURE 41. Explosion of large crystals of mercury fulminate initiated by impact. The first stage of explosion is a short lived rapid burning *AB* (350 m./sec.) followed by detonation *BC*.

FIGURE 43. Explosion of unconfined lead azide initiated by a hot wire at *A*. Only the detonation stage is observed.

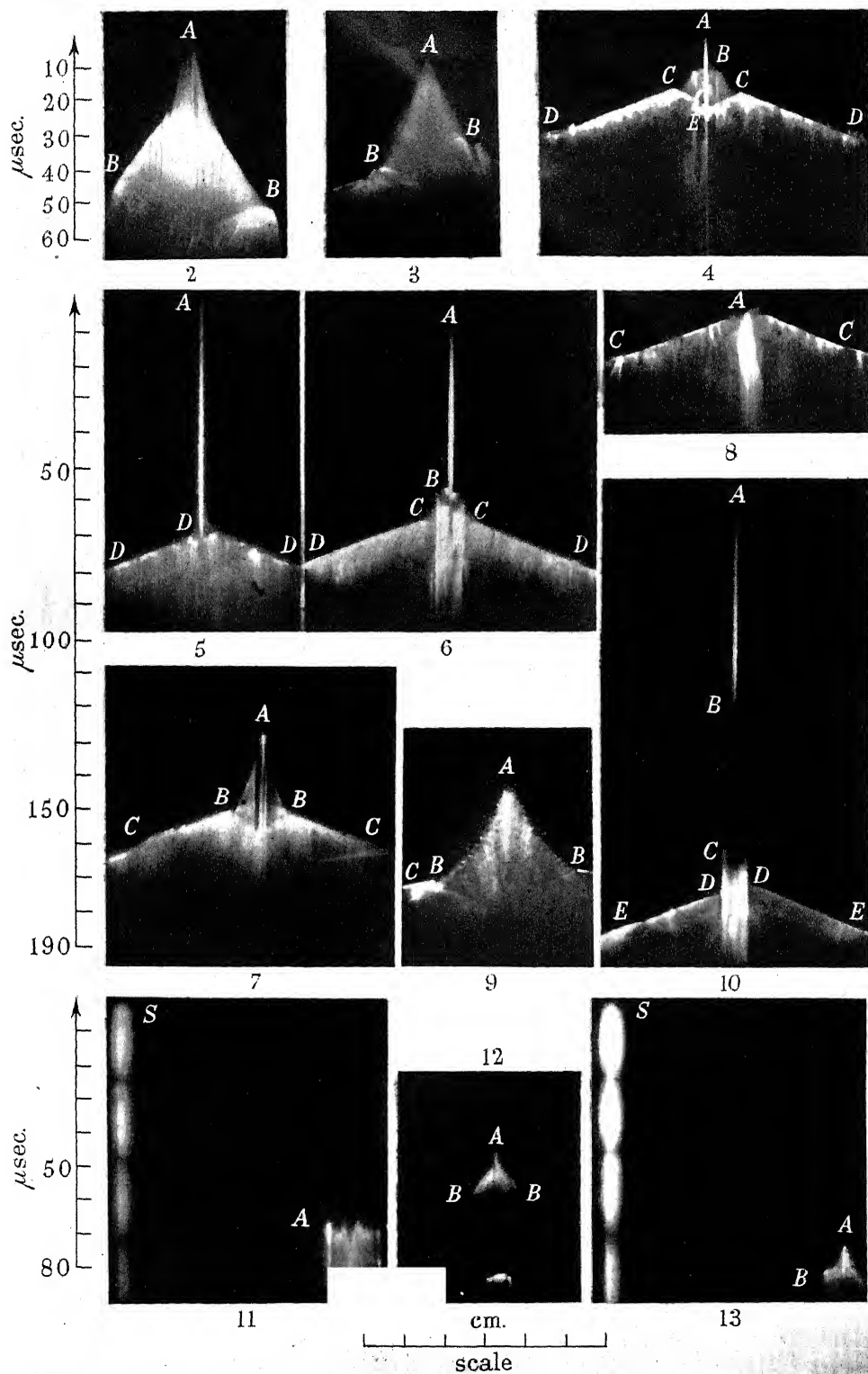
FIGURE 44. Explosion of a very thin layer of lead azide by a spark *A*. There is a short delay between the spark and the onset of detonation.

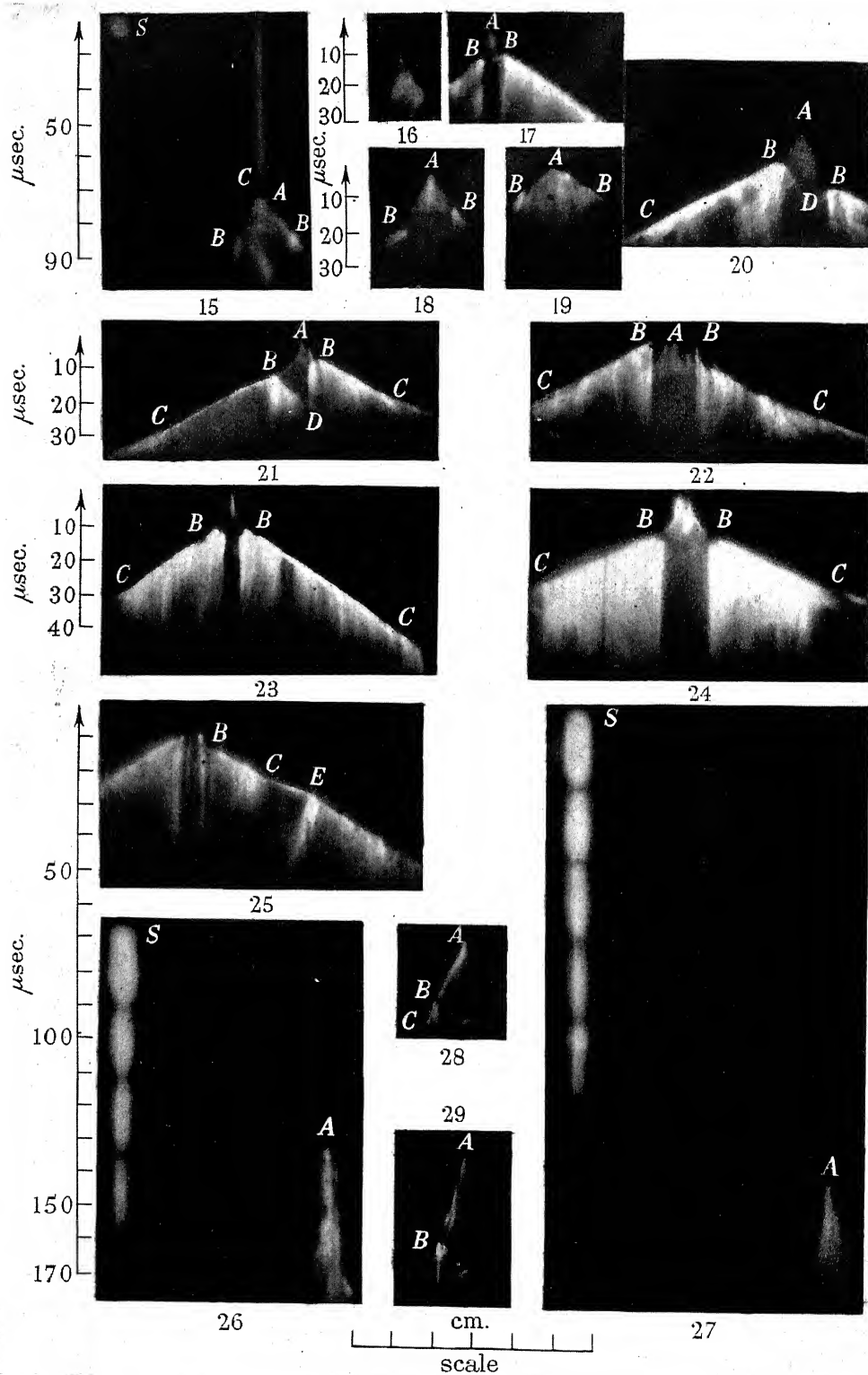
FIGURE 45. Explosion of mercury fulminate initiated by a spark *S*. A slow burning *SA* accelerates to a rapid burning *AB* from which a detonation *BC* begins at the discontinuity *B*.

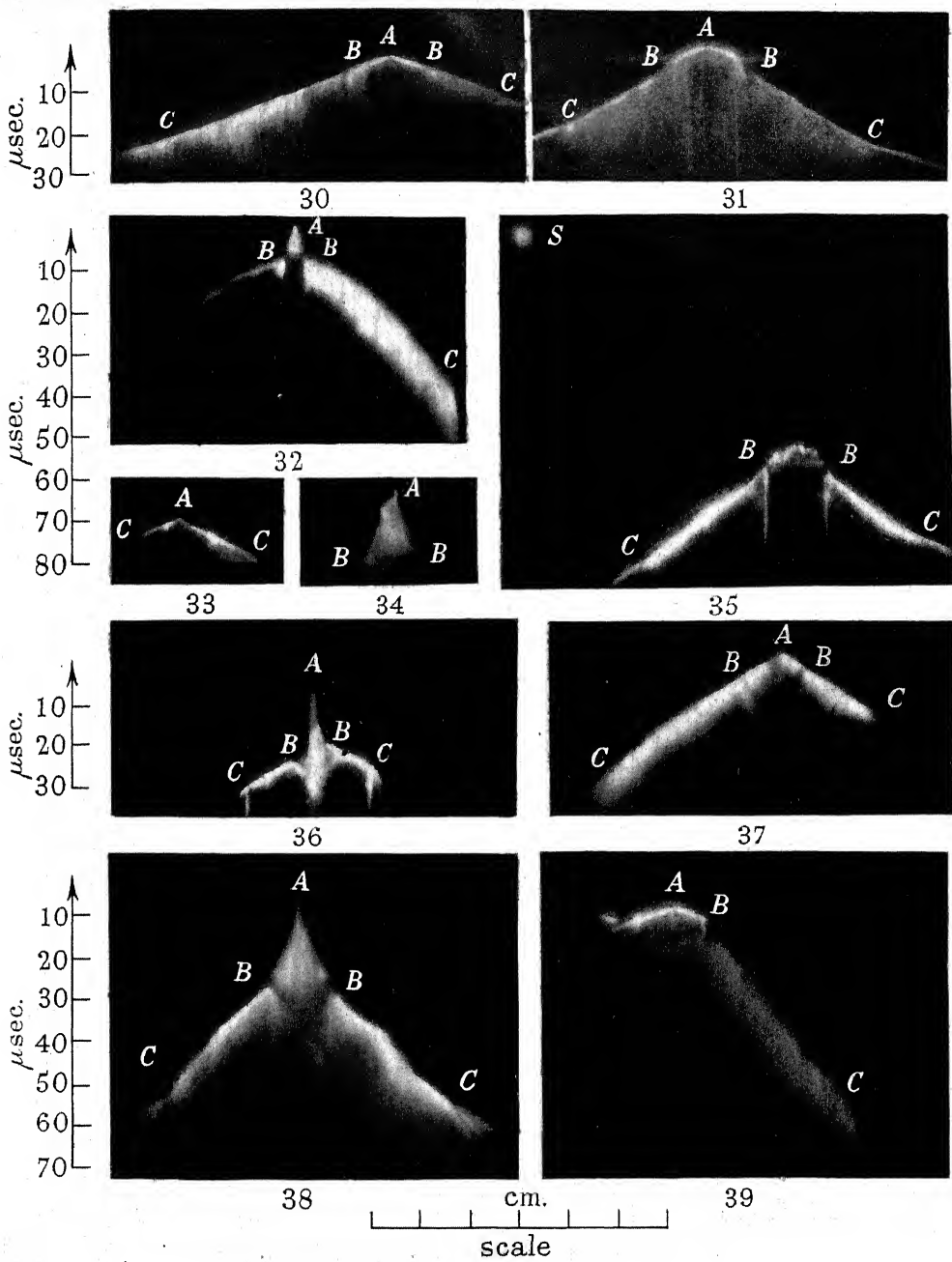
FIGURE 46. Similar to figure 45, but the onset of detonation occurs later, and not at a known mechanical discontinuity.

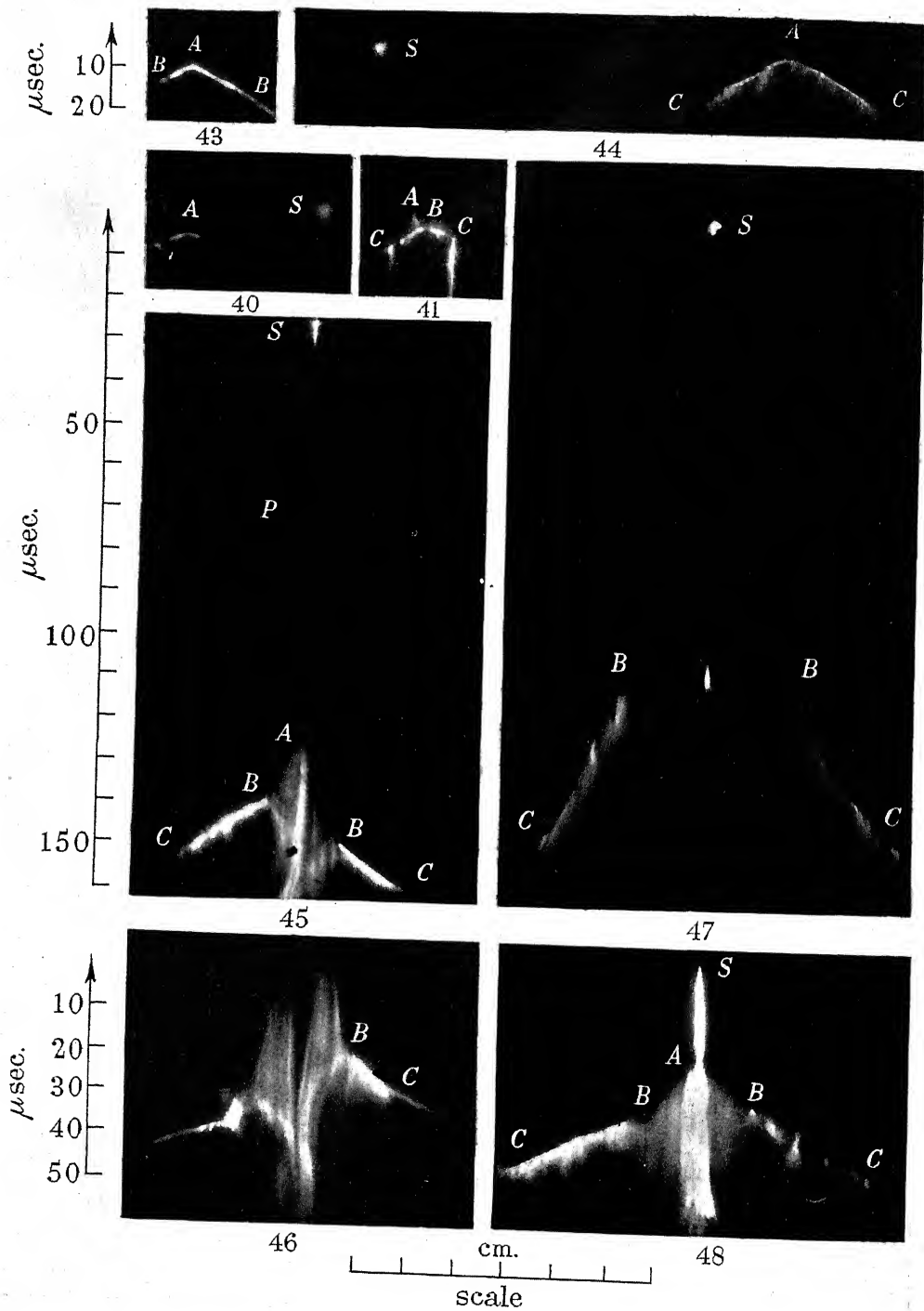
FIGURE 47. Initiation of lead styphnate by a spark *S*. The first part of the explosion is not recorded.

FIGURE 48. Explosion of P.E.T.N. initiated by a spark *S* while held under a heavy load. *AB* represents rapid burning in the constrained region. *BC* represents detonation.









Influence of entrapped gas on initiation of explosion in liquids and solids

By A. YOFFE

*Laboratory for the Physics and Chemistry of Rubbing Solids,
Department of Physical Chemistry, Cambridge*

(Communicated by F. P. Bowden, F.R.S.—Received 17 January 1949)

[Plate 15]

Further evidence has been obtained for the view that the initiation of explosion by impact in liquids is due to the compression and adiabatic heating of trapped gas. It is shown that the sensitivity of an explosive is very dependent upon the pressure ratio. Flat impact experiments on nitroglycerine spread as a ring on a flat anvil show that the explosion efficiency, which is high when the initial gas pressure is 1 atm., is reduced to zero when the initial gas pressure is c. 30 atm. A high explosion efficiency is still observed when the initial air pressure is less than 10^{-5} mm., and it is suggested that under these conditions the initiation is due to the compression of the nitroglycerine vapour itself present at a pressure of c. 10^{-8} mm. It is further suggested that the explosion begins in the vapour phase.

The behaviour of some solid secondary explosives such as P.E.T.N. and cyclonite shows many similarities to that of the liquid explosives. When spread as a ring rather than as a uniform film of crystals, an increase in explosion efficiency is again observed. When the impacts are carried out at an initial gas pressure of 100 atm. there is a considerable reduction in explosion efficiency. The striker must be sufficiently hard to make the explosive flow plastically so that gas pockets can be sealed off. Again it would appear that the adiabatic compression of trapped gas initiates the explosion. Even when spread as a continuous film of crystals, the air spaces which are present between the crystals may be sealed off by local melting or plastic flow during the impact and the adiabatic compression of these gas spaces may initiate the explosion.

INTRODUCTION

Previous work by Bowden, Mulcahy, Vines & Yoffe (1947) has shown that the sensitivity of liquid explosives to impact is greatly increased if small gas bubbles are trapped in the explosive during impact.

The initiation of explosion is of a thermal nature and is due to the adiabatic compression and heating during impact of the trapped bubbles. If p_1 is the initial pressure inside the gas bubble and T_1 its temperature, the final temperature T_2 reached inside the bubble when the final pressure has risen to p_2 is given by

$$T_2 = T_1 \left(\frac{p_2}{p_1} \right)^{\frac{\gamma-1}{\gamma}}. \quad (1)$$

The temperature rise $T_2 - T_1$ depends on the pressure ratio p_2/p_1 and on γ the ratio of the specific heats.

The main evidence which was advanced to support the view that initiation is due to the adiabatic heating of trapped gas was

- (i) the decrease in sensitivity when the small gas bubbles are eliminated,
- (ii) the decrease in explosion efficiency when gases are used which have a value for γ lower than that for air,

(iii) the very short time from impact to explosion and the location of the point of initiation at a compressed gas bubble.

This paper describes further investigations on the mechanism of initiation. The effects of the initial pressure and of the nature of the included gas have been studied more closely. In particular, if the incidence of explosion is related to T_2 of equation (1) then an increase in p_1 should lower the explosion efficiency. This point was investigated with an apparatus in which impact experiments could be carried out under pressures up to 100 atm. Another apparatus was constructed for carrying out impacts at pressures down to 10^{-6} mm. and which could also be filled with different gases.

A simple method of including the gas in the explosive was used: the liquid (or solid) is spread as a small ring on a flat anvil (see figure 1) and when this is struck with a flat hammer the small amount of gas in the centre is trapped and compressed. The initial volume of the gas is usually $c. 5 \times 10^{-5}$ ml. and regular explosions may be obtained with an impact energy of 300 to 500 g.cm.

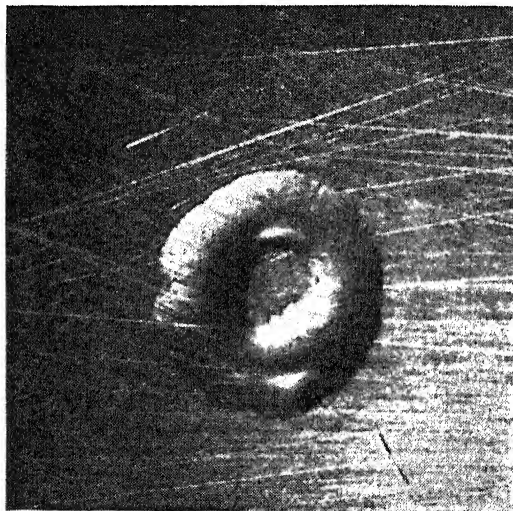


FIGURE 1. Nitroglycerine spread as a ring. (magn. $\times 2$.)

The investigations, hitherto confined to liquids, have been extended to solids. It has been found that the behaviour of such solid substances as P.E.T.N. and cyclonite is similar to that of nitroglycerine and other liquids. The paper is divided into two sections. The first deals with the initiation of explosion by impact in liquids, and the second describes initiation experiments with solid explosives.

I. LIQUIDS

Impact experiments at high initial gas pressures

Further proof that the initiation by gentle impact of explosion in nitroglycerine is due to the compression and adiabatic heating of trapped gas was obtained by carrying out experiments in a closed vessel in which the initial gas pressure p_1

could be varied between 1 and 100 atm. The final temperature T_2 will depend on the initial gas pressure p_1 . If explosion requires a definite high temperature T_2 to be reached, then initiation will be the more difficult the greater the initial gas pressure. The arrangement used was as follows. Nitroglycerine was spread as a ring on a flat steel anvil inside a chamber which was then filled with air or nitrogen from a gas cylinder. When the gas pressure had reached a required value, the explosive was hit with a flat steel hammer. A sketch of the apparatus is given in figure 2.

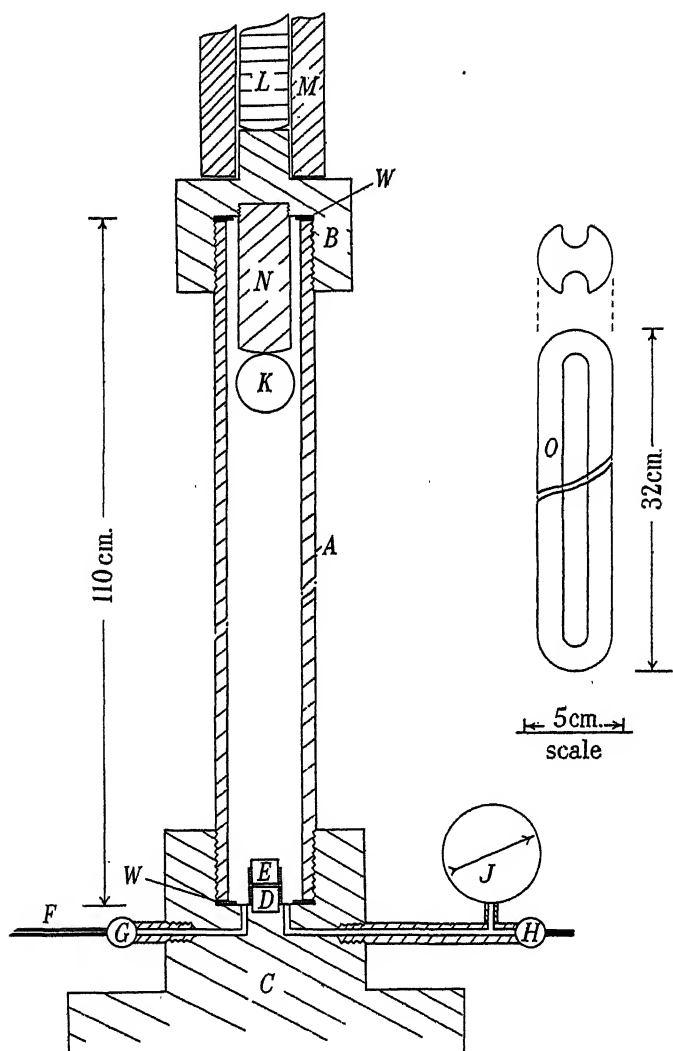


FIGURE 2. High pressure apparatus showing brass tube *A*; steel rollers *D*, *E*; steel ball *K*; electromagnet *LM*; steel head *B*; graphite washers *W*; gas inlet *F*, *G*; gauge *J* and outlet *H*. Cylindrical weight *O* shown on right.

Some results obtained at different initial gas pressures are summarized in table 1. When the steel ball is allowed to fall in the cylinder at high pressure a small correction has to be made to the energy of impact to allow for the viscous resistance of

the air. The correction is small and allowance has been made in the table for this effect.

The explosion efficiency falls as the initial gas pressure is increased and drops to zero when the initial pressure is 20 to 35 atm. The results provide strong evidence in support of the view that initiation is due to adiabatic compression of trapped gas. The difference between the results for air and nitrogen shows that the chemical nature of the gas is important. (See also Bowden *et al.* 1947; Mulcahy 1948).

TABLE 1. EXPLOSION EFFICIENCIES OBTAINED WITH A RING OF NITRO-GLYCERINE AT DIFFERENT INITIAL PRESSURES p_1 OF AIR AND NITROGEN

mass of striker (g.)	height of fall (cm.)	kinetic energy of impact (g.cm.)	initial gas pressure p_1 (atm.)	explosion efficiency	
				nitrogen (%)	air (%)
112	30	3.4×10^3	1	83	90
	36	—	10	30	—
	36	—	20	0	—
	36	—	25	—	20
	36	3.4×10^3	30	—	0
112	63.5	7.1×10^3	1	100	100
	98	—	20	10	—
	98	—	25	0	—
	98	—	30	—	36
	98	7.1×10^3	35	—	0

Impact experiments at low initial air pressures

The effect of reducing the initial air pressure below atmospheric has also been tried. Assuming the final pressure p_2 to be constant for a given impact energy, the final temperature T_2 increases as the initial pressure p_1 is reduced. Under these conditions, however, the mass of gas trapped in the bubble becomes smaller and the quantity of heat developed is correspondingly lowered.

The apparatus used is shown in figure 3 and consists of a collapsible bellows which was so designed that changes in gas pressure did not cause any relative movement of striker and anvil. The blow was delivered externally through a hardened surface. The nitroglycerine was spread as a ring on the anvil. With this apparatus the energy used in compressing the bellows itself is small. The pumping unit consisted of a Speedivac pump capable of pressures better than 10^{-4} mm. in series with a three stage mercury diffusion pump, and the pressures were recorded by a Pirani gauge having a range from 1 mm. to better than 10^{-4} mm. and a McLeod gauge calibrated to 10^{-6} mm.

Some results obtained when the initial air pressure was less than 10^{-5} mm. are given in table 2, and the explosion efficiency for one of the strikers has been plotted as a function of the height of fall in figure 4. In these experiments the total pressure did not drop much below 10^{-3} mm. which corresponds to the vapour pressure of nitroglycerine at room temperature. The energies of impact are such that no explosions are obtained when the nitroglycerine is spread as a continuous film.

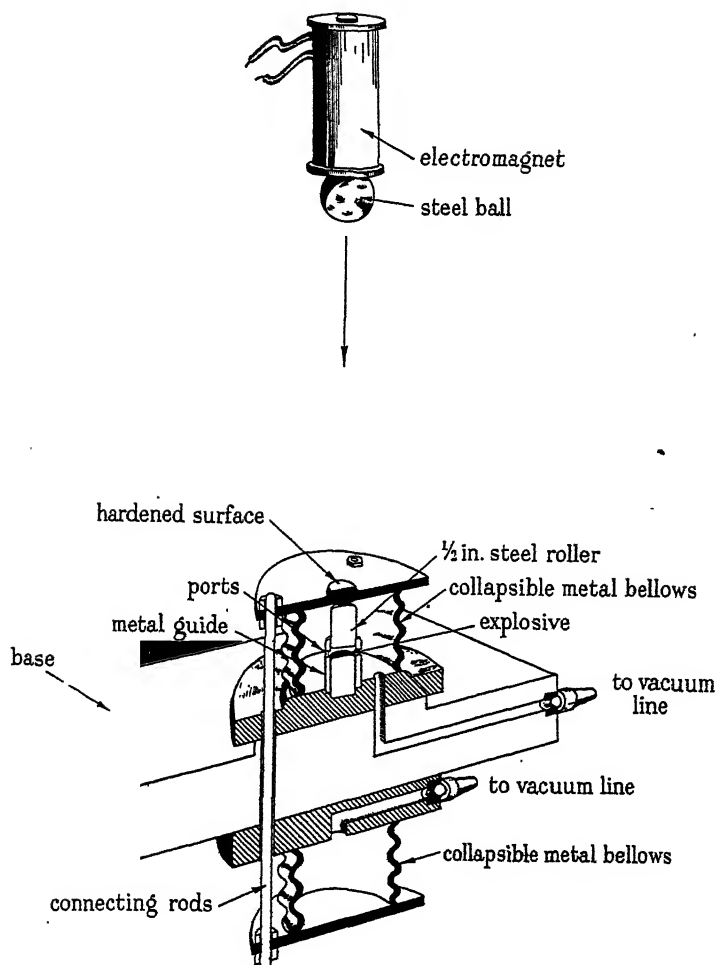


FIGURE 3. Low pressure apparatus (section through centre).

TABLE 2. EXPLOSION EFFICIENCY OBTAINED WITH A RING OF NITROGLYCERINE AT A LOW INITIAL AIR PRESSURE

mass of striker (g.)	height of fall (cm.)	energy of impact (g.cm.)	Initial pressure p_1 , total pressure c. 10^{-3} mm. air pressure c. 10^{-5} mm.		explosion efficiency (%)
			no. of explosions	no. of impacts	
210	30	6300	10/10		100
95	30	2850	23/26		88
95	10	950	8/14		57
95	5	480	1/12		8

The explosion efficiency for low energy impacts is appreciable even when the initial air pressure is less than 10^{-5} mm. The vapour pressure of the nitroglycerine itself is 10^{-3} mm.

Impact experiments in low pressure atmospheres of different gases

When gases having a value for γ lower than air are included in the gas space, there is a fall in explosion efficiency. Values have been obtained for the explosion efficiency over a fairly wide range of initial gas pressures. The nitroglycerine was spread as a ring and pure samples of the gases introduced into the apparatus described in the previous section.

The results obtained are given in figures 5 and 6. In figure 5 there is a comparison of the explosion efficiency when air $\gamma=1.4$, $n\text{C}_5\text{H}_{12}$ $\gamma=1.08$ and C_2H_4 $\gamma=1.26$ are used for an impact energy of 6.3×10^3 g.cm. It is seen that despite its smaller nominal value of γ C_5H_{12} vapour is more effective than C_2H_4 in enhancing the sensitivity to impact of nitroglycerine. Both are, however, less efficient than air, where the efficiency remains at 100 % even down to pressures of 10^{-5} mm. When the impact energy is 2.9×10^{-3} g.cm. the effect of C_5H_{12} and C_2H_4 is similar (see figure 6). In addition, low pressures of CCl_4 ($\gamma=1.13$) and CH_3ONO_2 were tried. The latter was chosen because of its similarity to nitroglycerine. The explosion efficiency remains unaltered when the evacuated space is filled with 50 mm. of methyl nitrate vapour. This result is of some interest from the practical point of view.

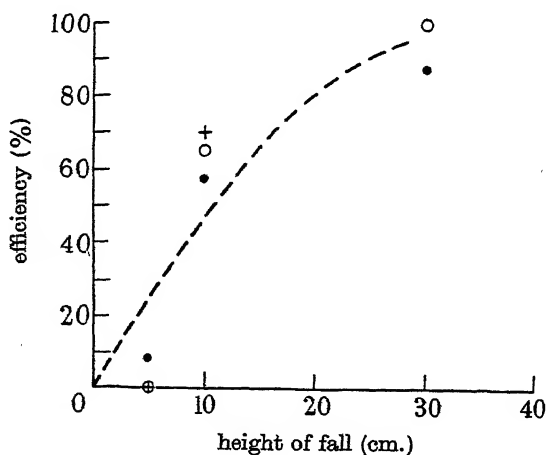


FIGURE 4. Explosion efficiency against height of fall curve obtained at low pressures of air and argon. Mass of striker 95 g. + argon, 20 mm. O air, 20 mm. • air, 10^{-5} mm.

DISCUSSION

The experiments described above demonstrate the sensitizing effect of bubbles of *n*-pentane, carbon tetrachloride and methyl nitrate. These are gases whose critical temperatures (470, 556 and c. 510° K) are above room temperature and which condense if compressed slowly. Initiation of explosion is facilitated even by high initial pressures, e.g. 250 mm. of *n*-pentane, that is under conditions where liquefaction could occur very readily during slow compression. The relative efficiencies (see figures 5 and 6) of these gases bear no relation to their several critical temperatures. It would appear that condensation does not occur during the very rapid (10^{-5} sec.) compression.

The adiabatic compression of saturated vapour need not always result in condensation. The condition for no liquefaction to occur is (Roberts 1943)

$$C_{\text{liq.}} - \frac{L}{T} + \frac{dL}{dT} < 0. \quad (2)$$

This relation is satisfied by carbon tetrachloride but not by *n*-pentane or methyl nitrate. However, if one takes the value $\gamma = 1.08$ for *n*-pentane, the estimated final

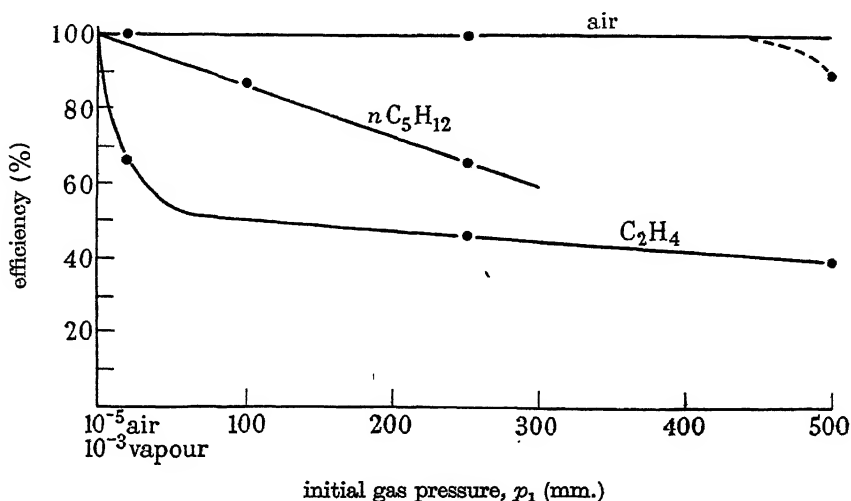


FIGURE 5. Explosion efficiency for different initial pressures of air, pentane and ethylene. Mass of striker 210 g. Height of fall 30 cm.

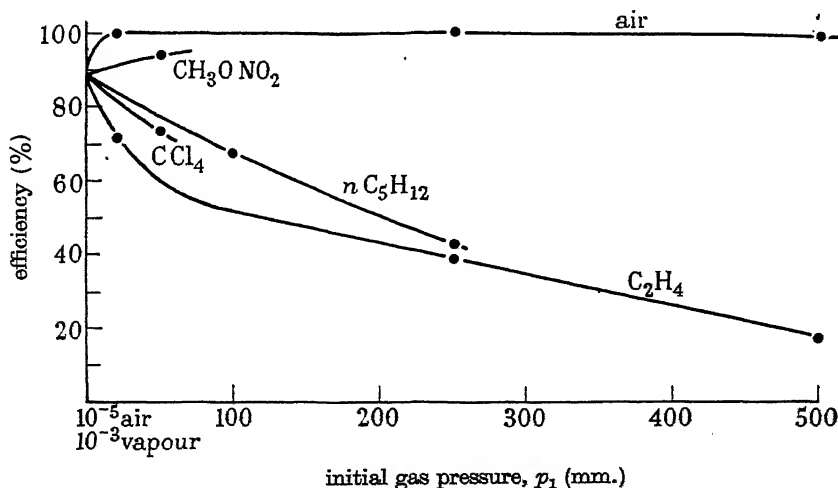


FIGURE 6. As for figure 5. Mass of striker 95 g.

temperature is *c.* 150° C which is much too low a temperature to explain the initiation of explosion (see Bowden & Gurton 1949, who quote 480° C as the ignition temperature).

The value of γ appropriate to the sudden compression of a gas and defined by equation (1), is not the value of γ at room temperature and pressure found from measurements of the speed of sound, etc. Even for the permanent gases, γ varies with pressure and temperature (see *International critical tables*, first edition) and during a compression γ depends on the rate of pressure rise (Lewis & von Elbe 1939). This variation has been ascribed to the failure of the kinetic energy to share itself out among all the modes of motion.

If n is the number of degrees of freedom involved,

$$\gamma = \frac{C_p}{C_v} = 1 + \frac{R}{n} \quad (3)$$

and any reduction in n causes an increase in γ towards the value $\gamma = 1.67$ corresponding to a monatomic gas. Experiments on the dispersion of sound in gases confirm this view but lead to numerically different results (Alexander & Lambert 1942; Richards 1939). It is possible that complex molecules such as pentane and methyl nitrate behave like monatomic gases during compression. However, the compression time before explosion occurs is fairly long (c. 10^{-5} sec.) compared with the period of relaxation during which equilibrium is established between translational and vibrational energy of the molecules. It should be remembered that γ as defined by equation (3) applies only to an ideal gas. This relation will hold for small compressions such as those occurring in a sound wave. When the pressure ratio is high deviations from the ideal nature of the gas must be taken into account. It can be shown that over a wide range of pressure, the deviation from ideal behaviour has a marked effect on the relation between temperature and pressure, and the actual value for the temperature rise is higher than that calculated from equation (1). This provides an explanation for the sensitizing action of the condensable gases and of the self-sensitized explosion of nitroglycerine spread as a ring *in vacuo* (10^{-5} mm. Hg). Here it is the vapour of the nitroglycerine itself which undergoes the adiabatic heating and which initiates the explosion.

Decomposition initiated in the vapour phase

The detailed mechanism of initiation cannot be inferred from these results alone, and the exact origin of decomposition is still uncertain. It may begin by the simultaneous (within 10^{-13} sec.) activation by hot molecules from the gas phase of two adjacent molecules in the interface (cf. Garner 1938). The exponential factor associated with this binary event is $e^{-2E/RT}$ and at the temperatures concerned (c. 500° C) is extremely small, and it is unlikely that this is the correct mechanism.

It is again suggested that the explosion begins as a burning in the gas phase of the vapour from the liquid explosive and that this inflammation spreads to the bulk, and the evidence for this view is summarized below:

1. The explosive vapour in the gas bubble is heated during compression to the same temperature as the gas. The liquid on the other hand must be heated by thermal conduction and it seems reasonable that the vapour will inflame first.
2. The concentration of nitroglycerine vapour in air at room temperature is extremely small (one part in 10^6). The possibility of the ignition of greatly diluted



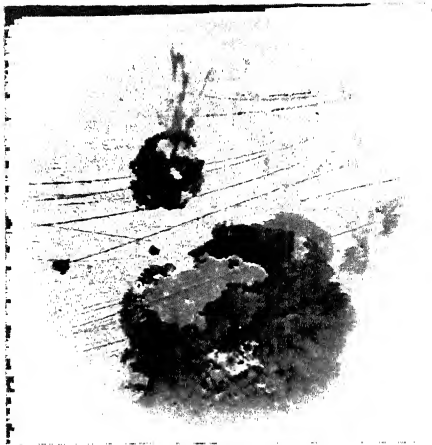
8a



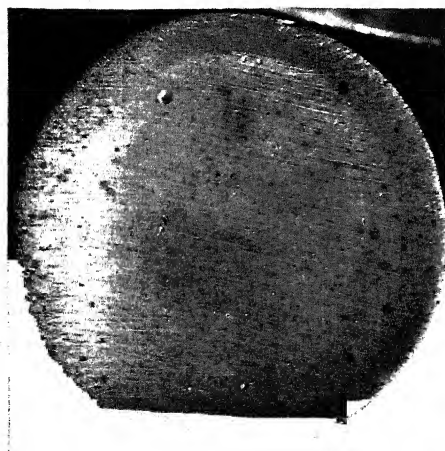
9



8b



10



8c



11

explosive vapour has not been closely examined. Preliminary experiments by Chamberlain, Gray & Walsh (1947) demonstrated the ease with which methyl nitrate vapour, diluted ten times with nitrogen, ignites, and confirm the likelihood of such inflammation.

3. When the initial air pressure is low (10^{-5} to 10^{-6} mm. Hg) the initiation is due to the compression and burning of the vapour of nitroglycerine heated to the ignition temperature.

4. Belajev (1938) has recorded the great difficulty of igniting, from the liquid phase, even such unstable liquids as nitrogen trichloride or nitroglycerine under high pressure, although estimated temperatures of c. 2000° C were produced.

5. The possibility of igniting gas mixtures by sudden compression is well known (e.g. in the diesel engine and the experiments of White & Price 1919, and Buckler & Norrish 1938).

6. The chemical nature of the included gas is important. Thus oxygen is more effective than nitrogen in sensitizing nitroglycerine to gentle impact (Mulcahy 1948). The results shown in figures 5 and 6 indicate that nitroglycerine and methyl nitrate vapour are more efficient than inert vapours such as *n*-pentane and carbon tetrachloride. The exothermic decomposition of methyl nitrate vapour greatly facilitates the growth of the explosion in the heated gas.

When liquid explosives are detonated by light impact, the following sequence of events is therefore postulated. Sudden compression and heating of the trapped gas occurs and exothermic decomposition of the explosive vapour begins. Even *in vacuo* the vapour of the explosive itself suffers a considerable temperature rise. Furthermore, the physical and chemical heating thus induced lead to evaporation of explosive from the walls of the bubble to give a richer mixture than existed before impact. Inflammation of the explosive vapour near the walls becomes sufficiently violent to ignite the liquid itself, and the explosion grows as a rapid burning through the bulk (cf. Bowden *et al.* 1947; Bowden & Gurton 1949).

II. SOLIDS

When a solid such as pentaerythritol tetranitrate (P.E.T.N.) is spread on the anvil as a ring and not as a uniform film of crystals its sensitivity to shock is increased.

The explosive (25 mg.) was placed between two hardened steel rollers, 1.3 cm. in diameter, and the blow delivered by a steel ball released from an electromagnet. The results obtained when the ball weighed 1860 g. are given in figure 7 where the explosion efficiency is plotted as a function of the height of fall. The energy necessary to give an explosion efficiency of 50 % is 2.8×10^4 g.cm. for a ring and 7.1×10^4 g.cm. for a continuous film.* This increase in sensitivity parallels that of liquid explosives although the effect is not so marked.

* In these experiments it was necessary to use dry P.E.T.N. otherwise low values were obtained for the explosion efficiency. This effect is difficult to explain since the presence of pentane or carbon tetrachloride had no desensitizing action (see later).

It could be argued that, in the case of a solid, the increase in efficiency with a ring of explosive is due to the greater pressure developed because of the smaller effective area of the strikers. However, though experiments with strikers of different areas have shown that the explosion efficiency does increase somewhat as the area of the striker is reduced, the effect is small compared with that observed with this particular distribution.

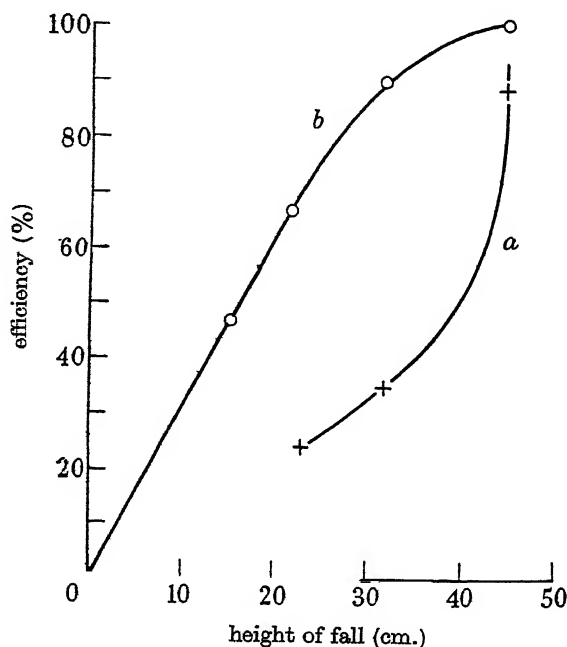


FIGURE 7. Explosion efficiency against height of fall curves for P.E.T.N. spread as a continuous film *a*, and as a ring *b*. Mass of striker 1860 g.

*Impact experiments in nitrogen at high pressure and
with strikers of different hardness*

The experiments with nitroglycerine at a high initial gas pressure confirmed the view that the initiation of explosion in liquids is due to the adiabatic compression of trapped gas. Similar experiments were therefore carried out with P.E.T.N.

TABLE 3. EXPLOSION EFFICIENCY OBTAINED WITH A RING OF P.E.T.N.
AT A HIGH INITIAL PRESSURE OF NITROGEN

height of fall (cm.)	mass of striker 2 kg. weight of sample 20 to 25 mg. hardness of rollers <i>D</i> and <i>E</i> 600 to 800 kg./mm. ²			explosion efficiency	
	initial gas pressure (atm.)	kinetic energy of impact (g.cm.)			%
23	1	4.6×10^4	7/12		58
25	1	5×10^4	10/10		100
36.5	100	5×10^4	2/15		13

The apparatus used is that shown in figure 2 except that the hardened steel weight *O* weighing 2 kg. was used instead of the spherical ball *K*.

The results with the P.E.T.N. spread as a ring are given in table 3. Measurements of the time of fall of the weight *O* showed that the viscous resistance of the air produced no appreciable reduction in velocity.

There is a reduction in the explosion efficiency from 100 to 13 % when the initial gas pressure is increased from 1 to 100 atm.

The effect of hardness of the striker

In carrying out impact experiments with solid secondary explosives, heavy strikers fall several feet and very high pressures may be developed in the solid, the limiting pressure being the dynamic flow pressure of the striker. When hardened steel strikers are used, the pressure in the P.E.T.N. film can reach 80,000 atm. before the steel deforms plastically. It is clear that the pressure ratio can still be very high even when the initial gas pressure is 100 atm. and the temperature rise in any trapped gas space can be considerable.

So as to reduce the maximum pressure attainable, strikers of metal softer than steel were also used. The results obtained in these experiments are summarized in table 4.

TABLE 4. EXPLOSION EFFICIENCIES

The sensitivity to shock of P.E.T.N. spread as a ring on a hard steel anvil under different initial gas pressures depends on the hardness of the striker. Mass of striker 2 kg.

height of fall (cm.)	striker	hardness of striker Vickers (kg./mm. ²)	initial gas pressure, nitrogen (atm.)	explosion efficiency
46	mild steel	200	1	3/3
56	mild steel	200	1	3/3
82	copper	72	1	0/5
	brass	120	1	0/3
	mild steel	200	1	6/6
	mild steel	200	100	2/10*

* These two explosions were of a very localized character.

No explosions are obtained with strikers of copper or brass. During the impact both the copper and brass strikers are deformed by the crystals of P.E.T.N. and a deep imprint of the explosive ring is left on the face of the striker (see figure 8, plate 15). The ring of P.E.T.N. appeared not to flow at all, and its dimensions were roughly the same before and after impact. When mild steel surfaces are used, explosions are readily obtained. Examination of the surface of the steel striker after explosion shows that only slight plastic deformation has occurred. Thus in order to initiate explosion it is necessary to make the P.E.T.N. flow, and for this to occur the hardness of the striker must be c. 200 kg./mm.². This means that the P.E.T.N. will be subjected to pressures of the order of 20,000 atm.

Mild steel strikers were also used at an initial gas pressure of 100 atm. The table shows that the explosion efficiency is lower, and the explosions obtained were very incomplete, the only evidence of decomposition being a small oxidation stain on the

steel surface. On the other hand the explosions at 1 atm. are generally complete and the efficiency when the same weight falls about half the height is about 100 %.

The difference between efficiencies at 1 and at 100 atm. pressure is striking. Even with nearly twice the height of fall the efficiency at 100 atm. is small and the explosions obtained are weak, despite the greater energy and momentum at the point of impact.

Impact experiments at low initial gas pressures

A series of experiments was carried out in a vacuum apparatus to see whether there was any change in explosion efficiency when the initial gas pressure was lowered. The apparatus used was similar to that shown in figure 3 but was more robustly made in order to withstand the high energy impacts. It could be evacuated at air pressures near 2×10^{-5} mm. and total pressure when the P.E.T.N. was present of c. 10^{-4} mm. The total pressure was measured by a Pirani gauge, which is a continuously recording instrument and which may detect a small amount of decomposition as soon as it occurs. The P.E.T.N. was spread as a ring on a flat steel roller and struck by another flat steel roller. The results obtained are given in table 5.

TABLE 5. EXPLOSION EFFICIENCY OF P.E.T.N. IN *VACUO* WHEN SPREAD AS A CONTINUOUS FILM AND AS A RING

Mass of striker 1860 g.
initial pressure of gas: 1. McLeod gauge: permanent gases 2 to 5×10^{-5} mm.
2. Pirani gauge: total pressure $< 10^{-4}$ mm.

height of fall (cm.)	explosion efficiency	
	continuous film	ring
80	6/9 ~ 66 %	2/2 —
45	4/12 ~ 33 %	14/18 78 %

The table shows that even at this low initial gas pressure the explosion efficiency is higher when the P.E.T.N. is spread as a ring. The explosion efficiency at the low gas pressure is lower than in air at 1 atm., the difference being more pronounced with the continuous film.

TABLE 6. IMPACT EXPERIMENTS ON A RING OF P.E.T.N. IN DIFFERENT GASES

Mass of striker 1860 g. Height of fall 45 cm.

gas	γ	initial gas pressure (mm. Hg)	explosion efficiency
air	1.4	760	8/8
ethylene	1.26	760	10/10
carbon tetrachloride	1.13	11	2/2
ether	1.08	300	5/5
n-pentane	1.08	300	3/3

The efficiency remains at 100 % with all these gases. This result is different from that observed with nitroglycerine where the efficiency decreased when gases of low γ were used. An explanation of this result is given in the discussion.

Impact experiments in different gases

Impact experiments on P.E.T.N. spread as a ring were carried out in gases having different values for γ . The gases used were air, ethylene, ether, *n*-pentane and carbon tetrachloride (see table 6).

The detonation of cyclonite by impact

Experiments similar to those described with P.E.T.N. were carried out with cyclotrimethylene trinitramine (cyclonite). The results follow the pattern described for P.E.T.N. though the difference in explosion efficiency of cyclonite when spread as a ring and as a continuous film is not so marked. Explosions could still be obtained when the initial air pressure was $\leq 2 \times 10^{-5}$ mm. and the pressure of cyclonite vapour $< 10^{-4}$ mm.

The propagating properties of cyclonite under impact conditions are poor. When spread as a continuous film, the explosions were nearly all partial. The extent of the explosion was much greater with the ring.

DISCUSSION

The behaviour of P.E.T.N. under impact shows many similarities to that of nitroglycerine and other liquid explosives. It is suggested that the initiation of explosion is of a thermal character and is due to the adiabatic compression of gas spaces trapped in the solid during the impact. The evidence for this conclusion is given below.

Several mechanisms exist by which the mechanical energy of the blow may be converted to heat. The energy of the impact may heat all the explosive uniformly. However, even assuming that all the energy is converted to heat, the temperature rise for P.E.T.N. is only 100°C when the impact energy is sufficient to initiate the explosion. The ignition temperature of P.E.T.N. when the time of heating is *c.* 10^{-4} sec. has been estimated at 400 to 500°C (see Bowden & Gurton 1949), so that the above temperature rise would be too small. Bernal (1938) proposed that the thermal energy could be concentrated at localized points in the explosive. Bowden and co-workers (1936, 1947) have shown that local hot spots could be produced during the rubbing of surfaces, the temperature rise being limited by the melting point of the solid. With materials such as P.E.T.N. and cyclonite the hot spot temperature due to frictional heating would not exceed about 140 and 200°C corresponding to the melting points, but again these figures are below the ignition temperatures.

Several workers (Rideal & Robertson 1948; Eirich (unpublished)) have postulated a mechanism for the initiation of explosion on the basis of the formation of a molten layer of the solid explosive. There is evidence that local melting of the solid does take place during impact, due either to frictional heating or to the application of a non-uniform pressure to the solid (*cf.* Johnston & Adams 1913). Mayes and Eirich (unpublished) observed melting of such materials as P.E.T.N. and sulphur during impact. Some of their experiments have been repeated, and the results may be seen in figures 9, 10 and 11, plate 15. Figure 9 shows the result of an

impact on P.E.T.N., and a fused (or flowed) layer may be seen on the anvil at the edge of the continuous film. The fused layer is most noticeable with low melting solids such as sulphur, figure 10. With waxy materials such as Tempilstik particles of molten substance are ejected from the impacting surfaces (figure 11).

Rideal and Robertson consider that the liquid explosive so formed is caused during impact to flow at a high speed through the small spaces between the explosive crystals and viscous heating raises the temperature of the liquid so much that rapid thermal decomposition occurs. This mechanism is not unlike that proposed for the initiation of thin continuous films of nitroglycerine in the absence of gas spaces (Bowden *et al.* 1947). However, it does not seem applicable to impact on solids under the usual conditions, since it is unlikely that that extreme velocity of flow is attained which the calculations of Cherry (1945) and Eirich & Tabor (1948) have shown to be necessary for the temperature rise due to viscous heating to become appreciable. It is more probable that the formation of a molten or plastic layer serves to seal off small air spaces, which—there is ample evidence—readily produce hot spots under impact. In other words, the melted and flowed explosive behaves just like a liquid explosive and it is reasonable to expect the mechanism of initiation to be the same as that already proved to be operative in the initiation of liquids.

The initiation of explosion by the sudden compression of gas spaces

When P.E.T.N. is spread as a ring and not as a continuous film the sensitiveness to impact is increased. During impact plastic flow or surface melting of the solid takes place and the relatively large gas space in the centre is trapped and compressed. The photographic investigations of Bowden & Gurton (1949) have shown that initiation begins at this trapped gas bubble. The formation of a hot spot near this gas pocket may be demonstrated by a simple experiment with red mercuric iodide. When spread as a ring and subjected to impact between flat surfaces, specks of the yellow modification appear close to the centre of the ring. These rapidly revert to the red form. The transition occurs at 126° C and does not occur on the slow application of a similar pressure.

Cyclonite and tetryl behave similarly to P.E.T.N. and sudden heating during adiabatic compression of the trapped gas appears to be the initiation mechanism for solid explosives of this type.*

A similar mechanism no doubt holds in the absence of a deliberately included gas space. Even when the crystals are spread as a 'continuous film' small gas spaces are trapped due to the plastic flow or surface melting of the crystals during impact.

The results with P.E.T.N. spread as a ring showed that at moderately low pressures gases such as carbon tetrachloride and *n*-pentane which have low values for γ have the same sensitizing effect on the explosion efficiency as the permanent

* The situation is evidently different with primary explosives which, in general, detonate before the melting point is reached. The formation of a molten layer is therefore excluded. Some experiments made with tetrazene suggest that the compression of trapped gas plays little part in the initiation. It seems probable in this case that the initiation is brought about by frictional heating as the crystals rub against each other or against the impacting surfaces (cf. Bowden & Gurton 1949).

gases. This result is different from that obtained with nitroglycerine where these gases were less efficient than air. The reason for this difference is not difficult to find. It has been shown that for the solid to flow under impact the pressure developed is much larger than that to be expected from impact on a liquid, and amounts to thousands of atmospheres. In P.E.T.N. it is in the region of 20,000 atm. and consequently the temperature rise due to adiabatic compression of gas spaces trapped in the solid will be very high even for gases of high specific heat. In the absence of any air when the initial pressure is low (c. 10^{-5} mm.) the space is filled with vapour of the explosive (c. 10^{-4} mm.) and once again it is the compression of the vapour which initiates the explosion.

The strongest evidence for the dependence of initiation on the sudden compression of trapped gas spaces is provided by experiments carried out at high gas pressures. It has already been shown that the maximum temperature attained in an adiabatic compression depends, not on the final pressure, but on the ratio of the final pressure to the initial pressure. Since the final pressure is in some way determined by the energy of the impact, it is clear that an increase in the initial gas pressure should reduce the maximum temperature attained in an included gas pocket. If initiation depends on the production of a local high temperature, then an initial high pressure would be expected to reduce the chances of an explosion if the impact conditions are unchanged. In the experiments with P.E.T.N. spread as a ring there was a marked drop in the explosion efficiency when the initial gas pressure was increased from 1 to 100 atm. of nitrogen.

It would seem therefore that there is little difference in the mechanism of initiation by impact in liquids, or solids such as P.E.T.N. and cyclonite. Under conditions where it is possible to include gas spaces in the explosive, initiation of explosion is due to the compression and heating of these trapped gas pockets.

I should like to thank Dr F. P. Bowden, F.R.S. for his constant interest and advice, the Royal Society for the Mackinnon Research Studentship, and the Ministry of Supply (Air) for grants for equipment. My thanks are also due to Mr P. Gray and Dr O. A. Gurton for helpful discussions.

REFERENCES

- Alexander, E. A. & Lambert, J. D. 1942 *Proc. Roy. Soc. A*, **179**, 499.
 Belajev, A. F. 1938 *C.R. Acad. Sci. U.R.S.S.* **18**, 267.
 Bernal, J. D. 1938 *Trans. Faraday Soc.* **34**, 1008.
 Bowden, F. P. & Gurton, O. A. 1949 *Proc. Roy. Soc. A*, **198**, 337.
 Bowden, F. P., Mulcahy, M. F. R., Vines, R. G. & Yoffe, A. 1947 *Proc. Roy. Soc. A*, **188**, 291, 311.
 Bowden, F. P. & Ridler, K. E. W. 1936 *Proc. Roy. Soc. A*, **154**, 640.
 Bowden, F. P., Stone, M. A. & Tudor, G. K. 1947 *Proc. Roy. Soc. A*, **188**, 329.
 Buckler, E. J. & Norrish, R. G. W. 1938 *Proc. Roy. Soc. A*, **167**, 292.
 Chamberlain, G. H., Gray, P. & Walsh, A. D. 1947 See P. Gray, *Conference on mechanism of inflammation and combustion processes in the gaseous phase*, Paris, 1948.
 Cherry, T. 1945 *Rep. Adv. Coun. Sci. Ind. Aust. A*, **121**, no. 9.
 Eirich, F. & Tabor, D. 1948 *Proc. Camb. Phil. Soc.* **44**, 566.

- Garner, W. E. 1938 *Trans. Faraday Soc.* **34**, 985, 1008.
 Jeffreys, H. 1935 *Phil. Mag.* **19**, 840.
 Johnson, J. & Adams, L. H. 1913 *Amer. J. Sci.* **35**, 205.
 Lewis, B. & von Elbe, G. 1939 *J. Chem. Phys.* **7**, 197.
 Mulcahy, M. F. R. 1948 *Phil. Mag.* **39**, 547.
 Richards, W. T. 1939 *Rev. Mod. Phys.* **11**, 36.
 Rideal, E. K. & Robertson, A. J. B. 1948 *Proc. Roy. Soc. A*, **195**, 135.
 Roberts, J. K. 1943 *Heat and thermodynamics*. Oxford University Press.
 White, A. G. & Price, F. W. 1919 *J. Chem. Soc.* p. 1462.

DESCRIPTION OF PLATE 15

FIGURE 8. (Magn. $\times 4.5$.) *a* plastic deformation of a copper striker under the region of the ring of P.E.T.N. after impact; *b* as *a* but using a brass striker; *c* mild steel striker showing little plastic deformation during impact and explosion.

FIGURE 9. Impact on P.E.T.N. crystals showing fused layer at the edge of the compressed powder. Striker 1860 g. falling 45 cm.

FIGURE 10. Melting of sulphur during impact.

FIGURE 11. Flow of Tempilstik (m.p. 316°C) during impact.

The effect of diffusion of the main reactants on flame speeds in gases

BY J. CORNER, *Armament Research Establishment, Ministry of Supply*

(Communicated by Sir John Lennard-Jones, F.R.S.—Received 15 January 1949—

Revised 7 April 1949)

The general equations of a single-reaction flame are written down, allowing for the diffusion of the main reactants and products. It is shown that the pressure-dependence of the flame speed is not altered by the consideration of diffusion. A method of successive approximation is applied to the equations with diffusion, giving an easily applied solution for the speed of a flame maintained by a single reaction whose rate depends only on the temperature and the concentrations of the reactants.

1. INTRODUCTION

Boys & Corner (1949) have discussed the speed of propagation of reaction zones (flames) when diffusion of the reactants can be neglected. In the present paper I shall discuss the effects produced by diffusion. The notation is, as far as possible, the same as in the earlier paper.

I begin (§ 2) by a sketch of the problem considered; in § 3 the equations of a moving reaction zone are written down; methods of dealing with these equations are reviewed in § 4, and an approximate solution, together with some general results, is given in § 5.

2. NOTATION; PROBLEM CONSIDERED

Let Ox be a fixed direction in space, normal to a plane front which is in *steady motion* in the direction of x decreasing. The unreacted gas is at rest at large negative values of x . One wishes to predict the speed for which such a steady motion is possible. Unsteady states or detonation, in which a shock wave is intimately associated with the reaction zone, are excluded, so that the equations to be written down do not explain how such a steady and stable flame could be generated. However, there are many cases in which a normal flame speed can be recognized, and to which the present work may apply. Practical application of the results has been made to a case of this kind (the burning of cordite).

'Flame speeds' mentioned in this paper are relative to the unburnt gas, being the 'fundamental flame velocity' of Coward & Hartwell (1932) or the 'burning velocity' of Lewis & von Elbe (1938a). For a discussion of definitions of flame speed reference may be made to Coward & Payman (1937).

As the flame passes through a layer of gas, the latter is heated, reacts and evolves heat. Since the pressure remains practically constant, the hot products of the reaction have a velocity (relative to the unburnt gas) in the opposite direction to the flame motion. Relative to the flame, the unburnt gas advances from the direction of negative x , and passes through the flame zone where there is a rapid but continuous rise of temperature and a continuous rise of the degree of reaction. The burnt gases, at the temperature of complete reaction, move along the x axis, with a velocity which is nearly always greater than the velocity of the unburnt gases on the other side of the flame. A complete theory of flame propagation should give the temperature and reaction distributions in the flame zone, and the velocity of the stable flame. This can be observed experimentally, and provides a test of the accuracy of the general picture of the reaction zone. The structure of the flame is not so easily accessible to experiment, except at low pressures.

Let the flame velocity be U_0 . To produce a stationary space distribution of temperature and composition the axes are given a velocity U_0 in the same direction as the flame. The velocity of the gas in the direction of x increasing is then U , a function of x alone. At large distances in the unburnt gas, U is U_0 . The choice of origin of x is arbitrary, and does not affect our equations.

The gas is assumed to undergo a single exothermic chemical reaction which maintains the flame. Let ϵ be the extent to which the reaction has proceeded when the gas has reached a plane x , so that of the gases passing this place a fraction ϵ (by mass) consists of the products of reaction.

As it stands, this assumption of a single reaction seems to be a serious restriction on the application of our equations. For there are few burnable gas mixtures in which several reactions do not go on together and in which reverse reactions can be neglected. A type of reaction to which our equations would apply would be a unimolecular decomposition with negligible reverse reactions. This system is particularly simple because of the small number of equations and functions involved. More general cases have more equations to determine a greater number of unknowns, but these equations are set up exactly as in the simple case. There are also many more

parameters such as diffusion coefficients of various substances and activation energies of various reactions, but there is no essentially new type as compared with the simple case considered here.*

In practical cases, it can often be assumed that certain reactions are fast compared with others, thus allowing the use of equilibrium constants instead of reaction rates, with a considerable simplification of some of the equations. However, the allowable simplifications vary so much from one case to another that it seems unlikely that any one theory can cover more than a fraction of the known flames. It is probable that each flame will have to be considered separately.

While the flame equations can be solved numerically whatever the reactions assumed, the labour is very great. Analytical solutions are desirable but seem to be unlikely except for simple models such as that of the present paper.

At a point x in the reaction zone, the temperature is $T^\circ \text{K}$, and the specific volume is $V \text{ cm.}^3/\text{g}$. One sq.cm. of the flame consumes unburnt gas at the rate $M \text{ g./sec}$. Since the motion is steady,

$$U = MV, \quad (1)$$

and M is the quantity to be found from the equations. Let suffixes 0, m refer to the initial and final states respectively, so that at $x = -\infty$, $T = T_0$, gas velocity = U_0 , specific volume = V_0 , and $\epsilon = 0$; at $x = +\infty$, the reaction has been completed, and $T = T_m$ (temperature of completely reacted gas), gas velocity = U_m , specific volume = V_m , and $\epsilon = 1$. Heat transmitted through the flame by radiation is neglected, as is cooling of the products, except for heat flow by conduction through the flame towards the cold reactant. This heat flow maintains the flame.

Usually U_m is larger than U_0 , so that, on transforming back to axes at rest with respect to the initial gas, the products are moving in the opposite direction to the flame. The alternative case could occur only if the flame were very cool and the number of molecules were sufficiently decreased in the flame reaction.

Let the diffusion coefficient of the reactant be $D \text{ cm.}^2/\text{sec.}$, the thermal conductivity of the mixture be $\lambda \text{ cal./cm. sec. deg. C.}$, and let the pressure be $P \text{ atm}$. In general, D and λ will depend on ϵ and the temperature.

Let $Q \text{ cal./g.}$ be the heat evolved by the reaction at temperature T , and let c'_p and $c_p \text{ cal./g.}$ be the constant-pressure specific heats of reactant and product at temperature T . We have

$$dQ/dT = c'_p - c_p. \quad (1a)$$

We need also the homogeneous reaction velocity. Let $\mathcal{R}(\epsilon, V, T)$ be the rate of change of ϵ when the reaction proceeds homogeneously at temperature T , with composition and specific volume given by ϵ and $V \text{ cm.}^3/\text{g}$. This reaction rate \mathcal{R} can be determined in principle by the methods of chemical kinetics, without experiments on flames.

The variation of any quantity θ , following the motion of an element of the fluid, is

$$D\theta/Dt = \partial\theta/\partial t + U\partial\theta/\partial x = U d\theta/dx. \quad (2)$$

* The work of Lewis & von Elbe (1934) on the theory of flames in ozone-oxygen mixtures shows the complications introduced by having even as few as three substances present, and the drastic simplifications necessary. The theory of composition variations in a reacting medium has, however, been formulated in a very general way by Damkohler (1936).

Therefore the rate at which an element of gas changes its composition by reaction is

$$[U d\epsilon/dx]_{\text{chem.}} = [De/Dt]_{\text{chem.}} = \mathcal{R}(\epsilon, V, T). \quad (3)$$

In the present model the diffusion of active particles (radicals or atoms) is not included. Only the diffusion of the main reactants and products is considered. The presence of active particles might be taken into account by generalizing the equations to apply to a multi-reaction system in which production, diffusion and decay of radicals, and the reactions they cause, are all included. The fact that the radicals are present only in small quantities simplifies the equations to some extent; for example, their contribution to the pressure can be neglected.

We have assumed that the reaction rate at any point depends on the temperature T at that point just as if the mixture were in a large vessel all at temperature T . This means that we are assuming that there is negligible temperature variation along a distance of the order of a mean free path in the gas. The condition can be put in the form: let P be the pressure in atm.; then the effective breadth of the flame must be much greater than $10^{-3}/P$ cm.

We take into account diffusion caused by concentration gradient, but neglect the relatively small 'thermal diffusion' due to the temperature not being uniform. The difference of composition set up by thermal diffusion between regions at 1000 and 2000° K is usually less than 7 %.

3. EQUATIONS OF THE REACTION ZONE

(a) Composition equation

Consider the region lying between x and $x + dx$, and of unit cross-section. The mass of initial substance entering at x is $M(1 - \epsilon)$ g./sec., and so the net amount leaving the region is

$$-M(d\epsilon/dx) dx \text{ g./sec.}$$

By equation (3) the loss of initial substance by chemical reaction in the region is

$$M\mathcal{R} dx/U \text{ g./sec.}$$

The concentration of reactant is $(1 - \epsilon)/V$ g./cm.³. Therefore the diffusional flow entering across the plane x is

$$(D/V) (d\epsilon/dx) \text{ g./sec.,}$$

and the net loss from the region by diffusion is

$$\frac{d}{dx} \left[\frac{D}{V} \frac{d\epsilon}{dx} \right] dx \text{ g./sec.}$$

Since the system is in a steady state,

$$\frac{M}{U} \mathcal{R}(\epsilon, V, T) + \frac{d}{dx} \left[\frac{D}{V} \frac{d\epsilon}{dx} \right] - M d\epsilon/dx = 0. \quad (4)$$

(b) Heat equation

A second equation is provided by the condition that there is no loss of heat by the region between x and $x + dx$.

The heat entering at x by conduction is $-\lambda dT/dx$ cal./sec., and the net loss of heat by conduction is

$$\frac{d}{dx} \left[-\lambda \frac{dT}{dx} \right] dx \text{ cal./sec.}$$

The heat evolved by chemical reaction in the layer is $MQ\mathcal{R}dx/U$ cal./sec. The net loss of heat by mass flow is

$$M \frac{d}{dx} [(c'_p(1-\epsilon) + c_p\epsilon) T] dx \text{ cal./sec.}$$

As there is no net loss of heat,

$$\frac{d}{dx} \left\{ \lambda \frac{dT}{dx} \right\} + MQ \frac{\mathcal{R}}{U} - M \frac{d}{dx} [(c'_p(1-\epsilon) + c_p\epsilon) T] = 0. \quad (5)$$

Elimination of \mathcal{R} from (4) and (5) gives

$$\frac{d}{dx} \left\{ \lambda \frac{dT}{dx} \right\} + Q \frac{d}{dx} \left(\frac{D}{V} \frac{d\epsilon}{dx} \right) - MQ \frac{d}{dx} (1-\epsilon) - M \frac{d}{dx} [(c'_p(1-\epsilon) + c_p\epsilon) T] = 0. \quad (6)$$

(c) *Momentum equation*

Conservation of momentum gives the hydrodynamical equation of motion: $P + MU$ independent of x . The pressure at any point is connected with the composition, concentration and temperature at that point. The relation may be written

$$P = f(\epsilon, V, T), \quad (7)$$

and by substituting this in the momentum equation one obtains a single relation between ϵ , V and T . But at all practical flame speeds P is effectively constant through the reaction zone, and with ample accuracy (7) alone gives the connexion between T , V and ϵ (taking P as the pressure in the undisturbed gas).

(d) *Boundary conditions*

The boundary conditions on these equations must suffice to determine ϵ , V and T as functions of x , and a solution must be possible only for one value of M . All the other quantities introduced are known from the nature of the mixture being burnt and its pressure in the undisturbed state. There are two differential equations (4) and (5), and one ordinary equation (7). The latter can be used to eliminate V in terms of ϵ , T and of course P , which is one of the given parameters of the problem. There remain two second-order differential equations for ϵ and T as functions of x .

For any value of M , the solution is determined if it must have known values of ϵ , $d\epsilon/dx$, T , dT/dx at a given point. The physical boundary conditions are that in the undisturbed gas there is no reaction and the temperature is T_0 ; that is, at $x = -\infty$, $T = T_0$, $\epsilon = 0$, $d\epsilon/dx = 0$, and therefore $dT/dx = 0$.

To form an acceptable picture of the reaction zone, there must be a point where the reaction has reached completion ($\epsilon = 1$), and at the same time the temperature has reached T_m , corresponding to complete reaction. These two conditions determine the point at which this happens and the value of M for which the condition is satisfied. That there is only one such flame speed (in steady motion) is suggested by experiment as well as made plausible by numerical solution of simple forms of the equations. If, on the other hand, pressure variations through the flame had been taken into account, with the possibility of shock waves, for example, then in some cases two widely different values of M would give solutions which satisfy all the boundary

conditions. The smaller M would be the ordinary 'flame speed', the larger M the 'detonation velocity'.

With most simple forms of reaction kinetics, the end-point of the reaction zone, where $\epsilon = 1$, is at infinity, and the boundary condition is then: T tends to T_m and ϵ to 1 as x tends to infinity. It is possible that in other cases the end-point may be at a finite distance, in which case the region of burnt gas would have $\epsilon = 1$ and $T = T_m$ for all x greater than the end-point.

4. FLAME THEORIES

The subject up to 1935-6 has been reviewed by Jost (1935, 1936) and Lewis & von Elbe (1938*b*). In this section only a sketch is attempted.

There are three equations of the problem, a reaction equation (4), a heat equation (5) and the equation of the isobars of the gas mixture (7). Most of the early theories used the heat equation (5), and avoided the reaction equation (4) by some plausible assumption, such as the introduction of an 'ignition temperature' or 'mean lifetime in the flame'. The problem was thus reduced from the solution of two simultaneous differential equations to the solution of a single equation. The earliest work on these lines was carried out by Mallard (1875) and Mallard & le Chatelier (1883), followed by Crussard (1914), Nusselt (1915) and Daniell (1930).

Jouguet (1913*a*) was the first to study the full set of flame equations (neglecting diffusion) as a system to determine the flame speed and the distribution of temperature and reaction through the flame zone. Jouguet explained how the flame speed arises from the need to satisfy all the boundary conditions. In a later paper, Jouguet (1913*b*) illustrated his general method by solving a special case, chosen for its mathematical simplicity, and which nevertheless retained the main features of a real flame. Some of the simplifications were removed later (Jouguet & Crussard 1919; Jouguet 1924).

Impressed by the role of active particles in promoting reaction, Lewis & von Elbe (1934) attempted the relatively simple multiple-reaction case of O_2 - O_3 mixtures. Drastic simplifications had to be made in the equation of conservation of energy, and were criticized later by Jost & Muffling (1937).

More recently the equations of the flame have been studied by Zeldowitsch & Frank-Kamenetsky (1938) with the aims of including diffusion of reactants and eliminating the assumption of an ignition temperature. Their equations differed in detail from those given in § 3. An approximate solution was obtained whose physical basis was that most of the reaction was assumed to take place near the maximum temperature T_m , on account of the rapid increase of reaction rate with temperature.

A more recent theory is that of Boys & Corner (1949), intended to apply to reactions just outside condensed phases. The reaction rate was taken to have a temperature-dependence $e^{-A/RT}$, and three cases were discussed: a first-order rate from a unimolecular mechanism; a second-order rate from a bimolecular reaction; a second-order rate as the low-pressure result of a unimolecular breakdown. Diffusion was neglected. The equations were solved by a method of successive approximation, whose accuracy was compared with a numerical solution. The first approximation,

of the same nature as that of Zeldowitsch & Frank-Kamenetsky, was wrong by a factor as large as three; the second approximation was correct to within 10 to 15 % in the examples studied.

5. SOLUTION OF THE FLAME EQUATIONS WITH DIFFUSION

(a) Combination of parameters

We shall examine first the results of combining the parameters of the equations, a method used by Boys & Corner (1949). We assume that $c_p = c'_p = c$, and that c , λ and D can be given mean values independent of T .*

The equations are

$$-D \frac{d}{dx} \left[\frac{1}{V} \frac{d\epsilon}{dx} \right] + M \frac{d\epsilon}{dx} = \frac{\mathcal{R}}{V}, \quad (8)$$

$$-\lambda d^2T/dx^2 + McdT/dx = Q\mathcal{R}/V. \quad (9)$$

We assume that the system is a mixture of perfect gases, the reactant having average molecular weight W , the products an average weight w . Let $n = W/w - 1$. Then

$$PVW/RT = 1 + n\epsilon. \quad (10)$$

The gas constant R depends on the units chosen for P and V ; if these are atmospheres and cubic centimetres, then $R = 82.06$.

For a first-order reaction (case I of Boys & Corner's paper)

$$\mathcal{R} = B_1(1 - \epsilon)e^{-A/RT}. \quad (11)$$

Let $y = x(PWB_1/\lambda)^{\frac{1}{2}}$; substituting this and (11) in (8) and (9),

$$M(PWB_1/\lambda)^{\frac{1}{2}} d\epsilon/dy - (PWB_1/\lambda) D \frac{d}{dy} \left\{ \frac{1}{V} \frac{d\epsilon}{dy} \right\} = PWB_1(1 - \epsilon)e^{-A/RT}/RT(1 + n\epsilon), \quad (12)$$

$$Mc(PWB_1/\lambda)^{\frac{1}{2}} dT/dy - PWB_1 d^2T/dy^2 = QPWB_1(1 - \epsilon)e^{-A/RT}/RT(1 + n\epsilon), \quad (13)$$

which can be written as

$$(DPW/\lambda) \frac{d}{dy} f_1(\epsilon, T, n, d\epsilon/dy) + M(\lambda PWB_1)^{-\frac{1}{2}} d\epsilon/dy = F(\epsilon, T, n, A), \quad (14)$$

$$Mc(\lambda PWB_1)^{-\frac{1}{2}} dT/dy - d^2T/dy^2 = QF(\epsilon, T, n, A), \quad (15)$$

where F and f_1 need not be written explicitly. The important point is that they depend only on the variables listed. The boundary conditions are:

- (i) $T = T_0$; $\epsilon = 0$; $d\epsilon/dy = dT/dy = 0$, at $y = -\infty$,
and (ii) $T = T_m$ where $\epsilon = 1$.

The boundary conditions and equations (14) and (15) contain the independent variable y , the dependent variables ϵ and T , and the parameters T_0 , T_m , n , DPW/λ , A , c , Q and $M(\lambda PWB_1)^{-\frac{1}{2}}$.

As T_0 , T_m , c and Q are related by a thermochemical equation, one of these four is redundant; we shall omit Q . There are no other quantities in the problem. The

* Of these, only D varies rapidly with T . Unfortunately, D behaves like T^2 or $T^{\frac{1}{2}}$.

equations and conditions can be satisfied simultaneously only if there is some relation between the parameters. We may write this relation as

$$M = (\lambda PW B_1)^{\frac{1}{2}} g(T_0, T_m, DPW/\lambda, n, A, c), \quad (16)$$

where g is a function only of the quantities listed. The corresponding result without diffusion, given by Boys & Corner, was that

$$M = (\lambda PW B_1)^{\frac{1}{2}} h(T_0, T_m, n, A, c).$$

DP/λ is independent of P except at very high pressures; hence the dependence of M on pressure is not altered by the consideration of diffusion.

$\{y\}$, the change of y between points at which T takes given values (T_1, T_2), is a function of these values and also of $M(\lambda PW B_1)^{-\frac{1}{2}}, T_0, T_m, n, A, c$ and DPW/λ . Hence $\{x\}$, the corresponding change in x , is

$$\{x\} = (\lambda/PW B_1)^{\frac{1}{2}} j(T_0, T_m, DPW/\lambda, n, A, c, T_1, T_2), \quad (17)$$

and in particular the effective flame thickness varies with pressure in the way predicted by the theory without diffusion.

The method can be applied to the other types of reaction discussed by Boys & Corner, namely,

Case II: $\mathcal{R} = B_2(1-\epsilon)^2 e^{-A/RT}/V$, a second-order reaction rate;

Case III: $\mathcal{R} = B_3(1-\epsilon)(1+n\epsilon) e^{-A/RT}/V$, a second-order rate which would be produced as the low-pressure form of a bimolecular reaction. For these,

$$M = PW(\lambda B)^{\frac{1}{2}} k(T_0, T_m, DPW/\lambda, n, A, c) \quad (18)$$

$$\text{and} \quad \{x\} = (\lambda/BP^2W^2)^{\frac{1}{2}} l(T_0, T_m, DPW/\lambda, n, A, c). \quad (19)$$

The functions k and l are different in cases II and III.

(b) Dependence on λ and D

It will be noticed that all the previous results have the form

$$M \text{ (or } \{x\}) = \lambda^{\frac{1}{2}} \times \text{function of } (D/\lambda).$$

This can be proved, by the methods of the preceding section, under more general conditions, which are that the conductivity and the diffusion coefficient D are independent of position in the flame.

Theories neglecting diffusion have given the result that the flame speed is proportional to $\lambda^{\frac{1}{2}}$, and this has been tested by experiment. The comparison has been usually not very encouraging (cf. Coward & Payman 1937; Lewis & von Elbe 1938*b*), since the flame speed does not appear to increase so fast as $\lambda^{\frac{1}{2}}$. This conclusion is not rigorous, because the flame speeds are measured for a series of mixtures of gases with very different conductivities (for example, H_2 and O_2), and the change in λ is guessed from the change in the proportions of the two gases. The conductivity of a mixture is apt to vary in a manner which one would not expect from the conductivities of its components.

That the flame speed varies less rapidly than $\lambda^{\frac{1}{2}}$ has been taken to mean that the reaction rate does not depend solely on the concentrations and the temperature, but

is influenced also by the presence of small quantities of active particles (atoms or radicals). However, failure of the law $M \propto \lambda^{\frac{1}{2}}$ may sometimes be due to D/λ not being the same for all the mixtures tested.

Damkohler (1940) has published an experimental and theoretical paper on the effect of turbulence on the bunsen flame, at Reynolds numbers up to 17,000. The effects noted are (a) alteration of shape, caused by turbulence with scale of irregularities greater than the thickness of the flame zero; (b) increase of flame speed, due to increase of heat conductivity by small-scale turbulence. The flame theory used by Damkohler was of the simple Mallard type, and the rate of burning was proportional to $\lambda^{\frac{1}{2}}$, since diffusion of the reactant was neglected. The experimental behaviour was in reasonable agreement with the semi-quantitative theory.

(c) *Boundary conditions*

Write $G = \epsilon - (D/MV) d\epsilon/dx,$ (20)

so that (4) and (5) become $dG/dx = \mathcal{R}/MV$ (21)

and $\frac{d}{dx} \left[McT - \lambda \frac{dT}{dx} - QMG \right] = 0.$ (22)

The condition at the hot end of the flame is $\epsilon = 1$ and $T = T_m$ simultaneously.

Although this condition is sufficient to determine a solution, it may be useful to mention the behaviour of certain other quantities at and near the hot end of the flame. Provided the reaction rate \mathcal{R} is not infinite anywhere, we deduce from (8) and (9) the continuity of $d\epsilon/dx$ and dT/dx . These are zero in the fully reacted gas, and hence also at the hot end-point of the flame. From (20) it follows that G tends to unity continuously as the hot end of the flame is approached.

(d) *Approximate solutions*

Boys & Corner (1949) have shown how the equations, neglecting diffusion, can be solved by a method of successive approximation. The same method will now be applied to the equations with diffusion. The method will first be applied to a first-order reaction, for which the equations are most easily handled. In this case

$$\mathcal{R} = B_1(1-\epsilon)e^{-A/RT}. \quad (23)$$

It is assumed that the specific heats of reactant and products are c cal./g. and that c , λ and D/λ can be assumed independent of temperature. In real gases, c increases slowly with temperature, and λ is proportional to $T^{\frac{1}{2}}$ very closely; D/λ is proportional to T , roughly. It will appear that the flame extends to an infinite distance on the hot side, so that dT/dx tends to zero as T/T_m and ϵ tend to 1. Hence

$$dT/dx = M\{c(T-T_m) + Q(1-G)\}/\lambda. \quad (24)$$

Also, from (10), $V = RT(1+\epsilon)/PW.$ (25)

From (20), (24) and (25),

$$G = \epsilon - DPW\{c(T-T_m) + Q(1-G)\}(d\epsilon/dT)/\lambda RT(1+\epsilon). \quad (26)$$

This can be turned into an explicit equation for G as a function of ϵ , T and $d\epsilon/dT$. It is more convenient, however, to use it to find $d\epsilon/dT$ from ϵ , G and T , giving

$$d\epsilon/dT = \lambda RT(1+n\epsilon)(\epsilon-G)/DPW\{c(T-T_m)+Q(1-G)\}. \quad (27)$$

Equations (24) to (27) apply to any form of reaction velocity. Using (24) and the reaction rate (23) in (21),

$$dG/dT = \lambda B_1(1-\epsilon)PW e^{-A/RT}/M^2RT(1+n\epsilon)\{c(T-T_m)+Q(1-G)\}. \quad (28)$$

We now find a first approximation, valid near the 'burnt' end, by keeping only the terms dominant in this region. Write

$$\left. \begin{aligned} M^2R/\lambda B_1PW &= \chi, & \lambda B_1PW e^{-A/RT_m}/M^2RT_m(1+n)c &= e^{-A/RT_m}\chi cT_m(1+n) = \beta, \\ \lambda RT_m(1+n)/cDPW &= \gamma, & c(T_m-T)/Q &= \xi, & 1-\epsilon &= \eta, & 1-G &= \zeta. \end{aligned} \right\} \quad (29)$$

ξ , η , ζ are all positive by definition, and tend to zero as the hot boundary of the flame is approached. The approximate equations are, from (28),

$$\frac{d\zeta}{d\xi} = \frac{\lambda B_1PW e^{-A/RT_m}\eta}{cM^2RT_m(1+n)(\zeta-\xi)} = \frac{\beta\eta}{\zeta-\xi}, \quad (30)$$

$$\text{and from (27)} \quad \frac{d\eta}{d\xi} = \frac{\lambda RT_m(1+n)(\zeta-\eta)}{cDPW(\zeta-\xi)} = \gamma \frac{\{\zeta-\eta\}}{\{\zeta-\xi\}}. \quad (31)$$

The solution of these equations, passing through $\xi = \eta = \zeta = 0$, is

$$\left. \begin{aligned} \eta &= \theta\xi, & \zeta &= \psi\xi, \\ \text{where} \quad \theta &= \frac{\gamma}{2\beta} \left[\left\{ 1 + \frac{4\beta}{\gamma} \right\}^{\frac{1}{2}} - 1 \right], & \psi &= 1 + \beta\theta. \end{aligned} \right\} \quad (32)$$

For a second approximation (32) is used to eliminate ϵ and T in the exact equation (28), and the resulting separable equation between G and T can be integrated. This is the same as the procedure adopted for the equations without diffusion. When $D = 0$, $\beta/\gamma = 0$, and therefore $\theta = 1$ and $\eta = \zeta$, as, of course, must happen in this case by definition. The equation between G and T is

$$\begin{aligned} dG/dT &= \lambda B_1PW e^{-A/RT} \theta(1-G)/M^2RT\{1+n-n\theta(1-G)\}\{Q(1-G)-Q(1-G)/\psi\} \\ &= \theta\psi e^{-A/RT}/\chi TQ(\psi-1)\{1+n-n\theta(1-G)\}. \end{aligned}$$

$$\begin{aligned} \text{Hence} \quad (1+n)(G-1) + \frac{n\theta}{2}(G-1)^2 &= \frac{1+\beta\theta}{\chi\beta Q} \int_{T_m}^T e^{-A/RT} \frac{dT}{T} \\ &\simeq \frac{1+\beta\theta}{\chi\beta Q} \left\{ \frac{RT_m}{A} \right\} (e^{-A/RT} - e^{-A/RT_m}). \end{aligned} \quad (33)$$

When $G = 0$, T is so small that $e^{-A/RT}$ is negligible. Hence the equation which determines the flame speed is

$$1+n-\frac{1}{2}n\theta = (1+\beta\theta)(RT_m/A)e^{-A/RT_m}/\chi\beta Q = (1+\beta\theta)(1+n)(RT_m/A)(cT_m/Q). \quad (34)$$

M is found from χ , which enters (34) through β and θ ; the latter is a function of β , which is determined by χ . Equation (34) could be turned into an explicit expression for χ , but this would be cumbersome. It is easy to find χ from (34) by a few trials.

When $D = 0$, (34) reduces to the result

$$\chi = e^{-A/RT_m} / cT_m \{ (1 + \frac{1}{2}n) (A/RT_m) (Q/cT_m) - 1 - n \}. \quad (35)$$

A numerical example will show the order of magnitude of the effect of diffusion. A typical case would be: a unimolecular decomposition of a substance of molecular weight 75, into 5 molecules of average molecular weight 15; initial temperature 300°K; final temperature 2500°K; specific heat = 0.4 cal./g.; heat conductivity 2×10^{-4} cal./cm. sec. deg. C.; both these are assumed constant through the flame. The activation energy A is assumed to be 30 kcal./mole.

Table 1 shows the variation of χ and $\chi^{\frac{1}{2}}$ with the value of DP . $\chi^{\frac{1}{2}}$ is proportional to the flame consumption M . To give an idea of the magnitude of the diffusion rates in the table, it may be noted that for N_2O - CO_2 diffusion (extrapolated to 2000°K), DP is about 5 cm.²/sec. Table 1 shows that for such a diffusion coefficient, the flame speed calculated by neglecting diffusion would be more than twice the proper value.

TABLE 1. EFFECT OF DIFFUSION ON A FIRST-ORDER FLAME

DP (atm. \times cm. ² /sec.)	χ	$\chi^{\frac{1}{2}}$
0	21.8×10^{-8}	4.67×10^{-4}
1.25	10.6×10^{-8}	3.26×10^{-4}
2.5	6.37×10^{-8}	2.52×10^{-4}
5	3.40×10^{-8}	1.84×10^{-4}
10	1.74×10^{-8}	1.32×10^{-4}

A numerical integration of a particular first-order flame is discussed in § 5 (e). It is shown that in this case the approximate solution gives a flame speed about 5 % too big.

The solution is almost the same for the type of reaction denoted by 'Case III' in the paper by Boys & Corner (1949):

$$\mathcal{R} = B_3(1 - \epsilon)(1 + n\epsilon)e^{-A/RT}/V. \quad (36)$$

This means that reaction takes place in sufficiently violent collisions of reactant with any type of molecule; this could be the low-pressure form of a unimolecular decomposition. Let

$$M^2 R^2 / \lambda B_3 P^2 W^2 = \chi, \quad e^{-A/RT_m} / \chi T_m^2 (1 + n) c = \beta, \quad \lambda RT_m (1 + n) / c DPW = \gamma. \quad (37)$$

The solution is given by (32) and (34), as for case I. Only the definitions of χ and β differ from those in (29).

A bimolecular reaction (case II) can be solved by the same methods, but not without a considerable amount of numerical computation. Three tables (2, 3 and 4) are given here, which enable flame speeds to be calculated in a few minutes. These tables are believed to cover most of the cases that will be encountered in practice. The method of construction is described in the appendix.

It will now be explained how to use these tables. The reaction rate is

$$\mathcal{R} = B_2(1 - \epsilon)^2 e^{-A/RT}/V. \quad (38)$$

Write

$$M^2 R^2 / \lambda B_2 P^2 W^2 = \chi, \quad e^{-A/RT_m} / c(1 + n)^2 T_m^2 \chi = \beta, \quad \lambda RT_m (1 + n) / c DPW = \gamma. \quad (39)$$

The definition of β is somewhat different from that in case III. The equation which determines the flame speed is

$$\gamma E_0/\beta - \{2n/(1+n)\} (\gamma/\beta)^2 \int_0^{E_0} N dE + (n/(1+n))^2 (\gamma/\beta)^3 \int_0^{E_0} N^2 dE = (cT_m/Q) (RT_m/A). \quad (40)$$

In tables 2, 3 and 4 are listed $\gamma E_0/\beta$, $(\gamma/\beta)^2 \int_0^{E_0} N dE$ and $(\gamma/\beta)^3 \int_0^{E_0} N^2 dE$ for certain values of γ and $\beta\gamma/(\gamma+15)$. This choice of variable makes the results for different γ more easily comparable.

To avoid interpolation of double entry tables, the flame speed should be found in the following way. From the given data, calculate γ and pick the nearest of the tabulated γ . With this γ and an assumed β/γ enter tables 2, 3 and 4, and calculate the left-hand side of (40). Repeat until $(cT_m/Q) (RT_m/A)$ has been bracketed, and find β/γ by interpolation. This gives β corresponding to the γ used. The true diffusion coefficient varies by a factor of as much as two in going through the effective region of the reaction zone, and therefore the γ used in entering the tables will often correspond to a value of D which is probably as good as that in the original data. If, however, it is required to find the flame speed corresponding exactly to a given non-tabulated γ , repeat with a different (tabulated) γ , and interpolate to find β corresponding to the true γ . This interpolation is easily carried out, since the tables are given at equal intervals of $\log \gamma$.

TABLE 2. $\gamma E_0/\beta$

$\frac{\beta\gamma}{\gamma+15}$	γ						
	0.5	1	2	4	8	16	∞
1	0.390	0.376	0.380	0.404	0.447	0.510	0.596
1.5	0.354	0.337	0.336	0.351	0.384	0.431	0.517
2	0.330	0.311	0.306	0.316	0.342	0.381	0.461
3	0.298	0.277	0.268	0.272	0.290	0.318	0.386
4	0.276	0.255	0.243	0.244	0.257	0.279	0.335
5	0.260	0.238	0.226	0.224	0.233	0.251	0.299
6	0.248	0.226	0.212	0.209	0.216	0.230	0.271
7	0.238	0.216	0.202	0.197	0.202	0.213	0.248
8	0.229	0.207	0.193	0.187	0.190	0.200	0.229
9	0.222	0.200	0.186	0.179	0.181	0.189	0.214
10	0.216	0.193	0.179	0.172	0.172	0.179	0.201
11	0.210	0.188	0.173	0.165	0.165	0.171	0.190
12	0.205	0.183	0.168	0.160	0.159	0.164	0.180
13	0.201	0.179	0.163	0.155	0.153	0.157	0.171
14	0.197	0.175	0.159	0.151	0.149	0.151	0.163
15	0.193	0.171	0.155	0.147	0.144	0.146	0.156
16	0.189	0.168	0.152	0.143	0.140	0.142	0.150

Table 5 shows the effect of diffusion on a flame with the following characteristics: second-order reaction (case II); $T_0 = 300^\circ \text{K}$; $T_m = 2500^\circ \text{K}$; $c = 0.4 \text{ cal./g.}$; $A = 30 \text{ kcal./mole}$; $W = 75$; $n = 4$; heat conductivity $2 \times 10^{-4} \text{ cal./cm. sec. deg. C}$. The effect of diffusion is of the same order of magnitude as for a flame with a first-order reaction.

TABLE 3. $(\gamma/\beta)^2 \int_0^{E_0} N dE$

(Values in brackets were found by interpolation)

$\frac{\beta\gamma}{\gamma+15}$	γ						
	0.5	1	2	4	8	16	∞
1	0.052	0.071	0.095	0.125	0.155	(0.180)	0.255
1.5	0.042	0.057	0.077	0.100	0.125	(0.147)	0.212
2	0.036	0.048	0.065	0.085	0.107	(0.127)	0.183
3	0.029	0.038	0.051	0.067	0.084	0.099	0.146
4	0.024	0.032	0.042	0.056	0.071	0.084	0.123
5	0.021	0.028	0.037	0.049	0.061	0.073	0.106
6	0.019	0.025	0.033	0.044	0.055	0.065	0.093
7	0.017	0.023	0.030	0.039	0.050	0.059	0.083
8	0.016	0.021	0.028	0.036	0.045	0.054	0.076
9	0.015	0.020	0.026	0.034	0.042	0.050	0.070
10	0.014	0.019	0.025	0.032	0.039	0.047	0.064
11	0.013	0.018	0.023	0.030	0.037	0.044	0.059
12	0.013	0.017	0.022	0.028	0.035	0.041	0.055
13	0.012	0.016	0.021	0.027	0.033	0.039	0.052
14	0.012	0.015	0.020	0.025	0.031	0.037	0.049
15	0.011	0.015	0.019	0.024	0.030	0.035	0.046
16	0.011	0.014	0.018	0.023	0.029	0.034	0.044

TABLE 4. $(\gamma/\beta)^3 \int_0^{E_0} N^2 dE$

(Values in brackets were found by interpolation.)

$\frac{\beta\gamma}{\gamma+15}$	γ						
	0.5	1	2	4	8	16	∞
1	0.009	0.018	0.033	0.055	0.081	(0.109)	0.159
1.5	0.006	0.013	0.024	0.041	0.062	(0.085)	0.130
2	0.005	0.010	0.019	0.033	0.050	(0.070)	0.111
3	0.004	0.007	0.013	0.023	0.037	0.051	0.087
4	0.003	0.005	0.011	0.018	0.029	0.041	0.072
5	0.002	0.005	0.009	0.015	0.025	0.035	0.061
6	0.002	0.004	0.007	0.013	0.021	0.030	0.054
7	0.002	0.003	0.006	0.011	0.018	0.027	0.048
8	0.001	0.003	0.006	0.010	0.016	0.024	0.043
9	0.001	0.003	0.005	0.009	0.015	0.022	0.039
10	0.001	0.002	0.005	0.008	0.014	0.020	0.036
11	0.001	0.002	0.004	0.008	0.013	0.018	0.033
12	0.001	0.002	0.004	0.007	0.012	0.017	0.031
13	0.001	0.002	0.004	0.007	0.011	0.016	0.029
14	0.001	0.002	0.003	0.006	0.010	0.015	0.027
15	0.001	0.002	0.003	0.006	0.009	0.014	0.026
16	0.001	0.002	0.003	0.005	0.009	0.013	0.025

TABLE 5. EFFECT OF DIFFUSION ON A SECOND-ORDER FLAME

DP (atm. \times cm. ² /sec.)	$10^{-7}\chi^{\dagger}$
0	3.16
0.85	2.28
1.71	1.77
3.42	1.12
6.84	0.63
13.68	0.33

The behaviour of ϵ near the cold side of the flame is not immediately obvious from this method of approximation. G of course decreases continually from 1 to 0 on passing through the flame from hot to cold side. At the cool end there is a large region in which there is little chemical reaction, and so $G = 0$. In this zone, ϵ and T are connected by the relation, derived from (27),

$$d\epsilon/dT = \lambda RT(1 + n\epsilon)\epsilon/DPWc(T - T_0),$$

of which the solutions are

$$\epsilon/(1 + n\epsilon) = E \exp\left\{\frac{qT}{T_0}\right\} (T - T_0)^q, \quad (41)$$

where E is an arbitrary constant, and $q = RT_0/DPWc$.

(e) Detailed structure of a special case

To test the method of approximation, a typical first-order flame has been solved exactly by numerical integration of the equation. The case taken was that already given as an example (table 1), with $DP = 2.5$ (cm.²/sec.) \times atm. The approximate method shows that $\chi = 6.36 \times 10^{-8}$. Numerical integration with a few trial values of χ showed that this flame would really have $\chi = 5.76 \times 10^{-8}$. In this case, therefore, the approximate method gives a flame speed about 5% too large. This error is of the same order of magnitude as in the case of negligible diffusion (Boys & Corner 1949).

The numerical integration was started at the hot boundary of the flame, with the first approximation as a guide to the behaviour of the solution in this region. The integration was carried out in steps of 50° K near the highest temperatures and then in steps of 100° C. down to 1000° K. Below this there was practically no chemical reaction, and the value of G at the lower boundary ($T = 300^\circ$ K) was easily estimated. For the true χ , G is zero here, and hence the proper χ was found by an interpolation of the G 's for various χ .

For the case of negligible diffusion, it had been found convenient to start the integration at the cold end. This was not possible in the present case, because near the cold end ϵ is determined by equation (40), in which there is an unknown constant; the particular solution needed is settled by the behaviour at higher temperatures where G is not zero. Therefore it is simpler to start at the hot boundary, even though the convergence of the numerical integration process is rather tricky here.

The detailed structure of the flame is shown in figures 1 and 2, in which ϵ and G are plotted against T and $x(B_1 P)^{\frac{1}{2}}$ respectively. G rises smoothly from 0 to 1, and below 1000° K is effectively zero. On the other hand, ϵ does not approach zero until close to 300° K, and therefore ϵ is needed in a temperature region where D is much smaller than at high temperatures. The ϵ curve of figures 1 and 2 has the following parts: above 1000° K, it has $DP = 2.5$; below 1000° K, curve I has $DP = 2.5$ and curve II has $DP = 0.2$. The latter is a reasonable average for 300 to 1000° K, if we assume that $DP = 2.5$ refers to a temperature in the neighbourhood of 2000° K. It can be seen that even with the smaller diffusion coefficient there is an appreciable diffusion of products ahead of the flame. The case of a diffusion coefficient which depends on temperature will be sufficiently clear from the two curves given.

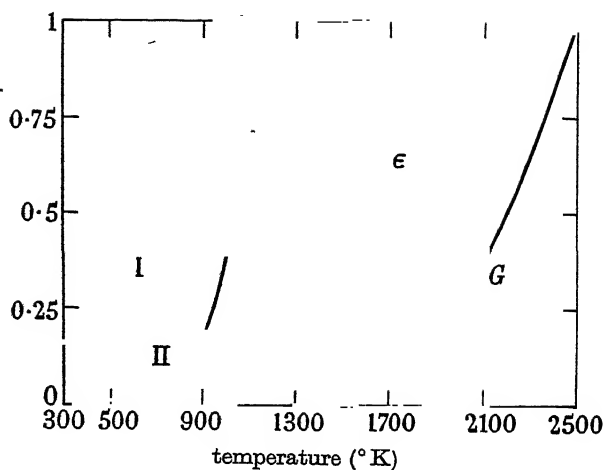


FIGURE 1. Structure of a typical first-order flame. Data in text.

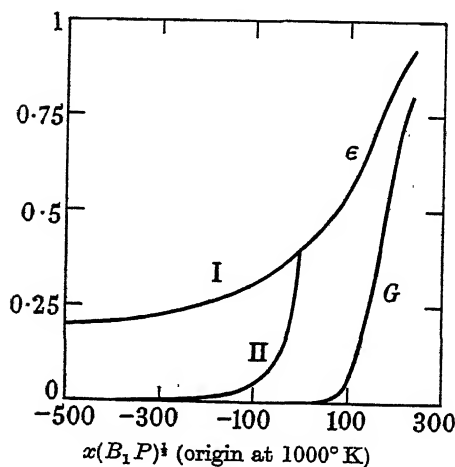


FIGURE 2. Structure of a typical first-order flame. Data in text.

APPENDIX

Approximate solution of flame with second-order reaction

$$\mathcal{R} = B_2(1-\epsilon)^2 e^{-A/RT}/V. \quad (42)$$

Write

$$\left. \begin{aligned} M^2 R^2 / \lambda B_2 P^2 W^2 &= \chi, & e^{-A/RT_m} / c(1+n)^2 T_m^2 \chi &= \beta, & \lambda R T_m (1+n) / c D P W &= \gamma, \\ (T_m - T) c / Q &= \xi, & 1 - \epsilon &= \eta, & 1 - G &= \zeta. \end{aligned} \right\} \quad (43)$$

The only difference from the notation of case III is in the definition of β . As usual

$$d\eta/d\xi = \gamma(\zeta - \eta)/(\zeta - \xi). \quad (44)$$

The exact equation for the reaction is

$$dG/dT = \lambda B_2 P^2 W^2 (1-\epsilon)^2 e^{-A/RT} / M^2 R^2 T^2 (1+n\epsilon)^2 \{c(T-T_m) + Q(1-G)\}, \quad (45)$$

and near the hot boundary this reduces to

$$d\xi/d\xi = \beta\eta^2/(\xi - \xi). \quad (46)$$

From (44) and (46), $d\eta/d\xi = \gamma(\xi - \eta)/\beta\eta^2$. (47)

Write $\xi - \eta = \gamma\Theta/\beta$ and $\eta = \gamma N/\beta$. Θ is positive for sufficiently small N , since $d\xi/dT$ is positive for T sufficiently close to T_m . (47) becomes

$$1 + d\Theta/dN = N^2/\Theta. \quad (48)$$

The solution required is that with $\Theta = 0$ for $N = 0$, and Θ positive for small N . This solution behaves as $\Theta = N^2$ for small N . Put

$$\Theta = N^2 f(N). \quad (49)$$

Then $1 + 2Nf + N^2 df/dN = 1/f$. (50)

As N tends to zero, f tends to unity. There is no useful solution in the form of a power series in N near $N = 0$. Numerical integration is also impracticable, at any rate in the range of N between 0 and 0.5, because $N^2 df/dN$ is obtained as the difference of two nearly equal quantities. This fact assures the rapid convergence of a process suggested by Dr A. F. Devonshire: $N^2 df/dN$ is neglected, and (50) solved for f as a function of N ; this first approximation f_1 is used in the $N^2 df/dN$ term, and the quadratic for f_2 solved. This process can be taken to the third approximation quite easily. Table 6 shows that the convergence is rapid even for N as large as 10. For the present purpose it is sufficient to use the third approximation.

TABLE 6. SUCCESSIVE APPROXIMATIONS TO f

N	f_1	$f_2 - f_1$	f_2	$f_3 - f_2$	f_3
0.25	0.7321	166	0.7487	-10	0.7477
0.5	0.6181	242	0.6423	-14	0.6409
2	0.3904	293	0.4197	4	0.4201
10	0.2000	208	0.2208	16	0.2224

The formulae are $f_n = \{[(1 + K_n)^2 + 8N]^{\frac{1}{2}} - (1 + K_n)\}/4N$,

where $K_1 = 0$, $K_2 = \frac{1}{4} \left[1 - \frac{1 + 4N}{(1 + 8N)^{\frac{1}{2}}} \right]$, and

$$K_3 = \frac{1 + K_2}{4} - \frac{\{(1 + K_2)^2 + 8N\}^{\frac{1}{2}}}{4} - \frac{N^2(1 + K_2)}{(1 + 8N)^{\frac{1}{2}} \{(1 + K_2)^2 + 8N\}^{\frac{1}{2}}} + \frac{N}{\{(1 + K_2)^2 + 8N\}^{\frac{1}{2}}} + \frac{N^2}{(1 + 8N)^{\frac{1}{2}}}.$$

The following asymptotic formula (suggested by Mr E. P. Hicks) was used for N greater than 8:

$$\Theta = 0.8164966N^{\frac{1}{2}} - 0.4N + 0.146969N^{\frac{3}{2}} - 0.048 + 0.022045N^{-\frac{1}{2}} - 0.014109N^{-1} + \dots \quad (51)$$

Write $\xi = \gamma E/\beta$. Substituting this and $\zeta = \gamma[N + N^2 f(N)]/\beta$ in (46),

$$N + N^2 f - E = \gamma N^2 f dE/dN. \quad (52)$$

The solution is needed which has $E = 0$ at $N = 0$. Solutions for certain special values of γ are: $\gamma = 1$, $E = N$; $\gamma = 0$, $E = N + N^2 f$; $\gamma = \infty$, $E = 0$. In the general case, write

$$E = N - \psi(N). \quad (53)$$

Then

$$d\psi/dN + \psi/\gamma N^2 f = (\gamma - 1)/\gamma. \quad (54)$$

Let $\int_1^N dN/N^2 f(N) = \phi(N)$. The required solution of (54) is

$$\psi(N) = \{(\gamma - 1)/\gamma\} e^{-\phi(N)/\gamma} \int_0^N e^{\phi(x)/\gamma} dx. \quad (55)$$

For small and large γ the computations are more easily carried out with slightly different forms of solution.

Using the table of $f(N)$, the function E was tabulated for various values of γ and N .

The second approximation is now constructed in the usual way. The equation connecting G and T is

$$\begin{aligned} dG/dT &= (1 - \epsilon)^2 e^{-A/RT} / \chi T^2 (1 + n\epsilon)^2 \{c(T - T_m) + Q(1 - G)\} \\ &= \eta^2 e^{-A/RT} / \chi T^2 Q(\zeta - \xi) (1 + n - n\eta)^2. \end{aligned}$$

Hence

$$\int_{T_0}^{T_m} \frac{e^{-A/RT}}{\chi Q T^2} dT = \int_0^1 \frac{(1 + n - n\eta)^2 (\zeta - \xi)}{\eta^2} dG. \quad (56)$$

The left-hand side $= R e^{-A/RT_m} / Q A \chi$, since e^{-A/RT_0} is small compared to e^{-A/RT_m} . The right-hand side of (56) is

$$\int_0^1 (1 + n - n\gamma N/\beta)^2 \beta \frac{d\xi}{d\zeta} d\zeta = \gamma \int_0^{E_0} (1 + n - n\gamma N/\beta)^2 dE,$$

where $E = E_0$ when $\zeta = 1$, that is, when

$$N + N^2 f(N) = \beta/\gamma. \quad (57)$$

Let N_0 be the solution of this equation. Hence

$$\gamma(1 + n)^2 E_0 - 2n(1 + n) \frac{\gamma^2}{\beta} \int_0^{E_0} N dE + \frac{n^2 \gamma^3}{\beta^2} \int_0^{E_0} N^2 dE = R e^{-A/RT_m} / Q A \chi,$$

which reduces to

$$\frac{\gamma E_0}{\beta} - \frac{2n}{1 + n} \left\{ \frac{\gamma}{\beta} \right\}^2 \int_0^{E_0} N dE + \left\{ \frac{n}{1 + n} \right\}^2 \left\{ \frac{\gamma}{\beta} \right\}^3 \int_0^{E_0} N^2 dE = (cT_m/Q) (RT_m/A). \quad (58)$$

This equation determines β and so M .

The upper limit E_0 is found by solving (57) for N as a function of β/γ , and hence finding E_0 as a function of γ and β/γ . Integrals such as $\int_0^{E_0} N dE$ are found by integration by parts; for example, $\int_0^{E_0} N dE = N_0(E_0 - N_0/2) + \int_0^{N_0} \psi dN$. For these purposes, $\int \psi dN$ and $\int N \psi dN$ were computed.

In § 5 (d) have been tabulated $\gamma E_0/\beta$, $(\gamma/\beta)^2 \int_0^{E_0} N dE$, and $(\gamma/\beta)^3 \int_0^{E_0} N^2 dE$ (tables 2, 3 and 4 respectively). They are functions of β and γ only.

The auxiliary functions $f(N)$ and $E(N, \gamma)$ do not appear to have any other application, so they are not given here.

For zero diffusion, tables have been given previously (Boys & Corner 1949). The functions $\gamma E_0/\beta$, $(\gamma/\beta)^2 \int_0^{E_0} N dE$ and $(\gamma/\beta)^3 \int_0^{E_0} N^2 dE$ for the limit of zero D (infinite γ) were tabulated in that paper as $g_1(1/\beta)$, $g_2(1/\beta)$ and $g_3(1/\beta)$. They have been incorporated in the present tables.

Certain results for large γ and small $\beta\gamma/(\gamma+15)$ were found by interpolation, to avoid a lengthy re-computation of certain functions at smaller intervals. These results are given in brackets.

I am indebted to Dr S. F. Boys for many discussions, to Professor Sir John Lennard-Jones, F.R.S., for his continued encouragement of this research, and to the Chief Scientist, Ministry of Supply, for permission to publish this paper.

REFERENCES

- Boys, S. F. & Corner, J. 1949 *Proc. Roy. Soc. A*, **197**, 90.
 Coward, H. F. & Hartwell, F. J. 1932 *J. Chem. Soc.* pp. 1996, 2676.
 Coward, H. F. & Payman, W. 1937 *Chem. Rev.* **21**, 259.
 Crussard, L. 1914 *C.R. Acad. Sci., Paris*, **158**, 125, 340.
 Damkohler, G. 1936 *Z. Elektrochem.* **42**, 846.
 Damkohler, G. 1940 *Z. Elektrochem.* **46**, 601.
 Daniell, P. J. 1930 *Proc. Roy. Soc. A*, **126**, 393.
 Jost, W. 1935 *Z. Elektrochem.* **41**, 183, 232.
 Jost, W. 1936 *Z. Elektrochem.* **42**, 461.
 Jost, W. & Muffing, L. V. 1937 *Z. phys. Chem. A*, **181**, 208.
 Jouguet, E. 1913a *C.R. Acad. Sci., Paris*, **156**, 872.
 Jouguet, E. 1913b *C.R. Acad. Sci., Paris*, **156**, 1058.
 Jouguet, E. 1924 *C.R. Acad. Sci., Paris*, **179**, 454.
 Jouguet, E. & Crussard, L. 1919 *C.R. Acad. Sci., Paris*, **168**, 820.
 Lewis, B. & von Elbe, G. 1934 *J. Chem. Phys.* **2**, 537.
 Lewis, B. & von Elbe, G. 1938a *Combustion, flames and explosions of gases*, p. 160. Cambridge University Press.
 Lewis, B. & von Elbe, G. 1938b *Combustion, flames and explosions of gases*, ch. XII. Cambridge University Press.
 Mallard, E. 1875 *Ann. Min., Paris*, **7**, 355.
 Mallard, E. & le Chatelier, H. 1883 *Ann. Min., Paris*, **4**, 274.
 Nusselt, W. 1915 *Z. Ver. deutsch. Ing.* **59**, 872.
 Zeldowitsch, J. & Frank-Kamenetsky, D. A. 1938 *C.R. Acad. Sci. U.R.S.S.* **19**, 693; *Acta Physicochim.* **9**, 341.

Paramagnetic resonance in the copper Tutton salts

By B. BLEANEY, R. P. PENROSE* AND BETTY I. PLUMPTON

Clarendon Laboratory, University of Oxford

(Communicated by F. Simon, F.R.S.—Received 27 January 1949)

Part I. Seven of the double sulphates and selenates ('Tutton salts') of copper have been examined by the method of paramagnetic resonance. The deviation of the effective gyromagnetic ratio from the free-spin value of 2 is shown to be generally consistent with the theory of Polder (1942), who assumes a crystalline electric field of tetragonal symmetry to act on each ion. In some salts an appreciable departure from tetragonal symmetry is observed.

Part II. The variation of the width of the absorption lines with the direction of the applied magnetic field is studied and compared with that calculated from magnetic dipole interaction. The line shapes show that exchange interaction is present in varying degree, though much less powerful than in copper sulphate. The 'mean square moment' of the lines can be explained by the magnetic dipole interaction, together with a contribution from a hyperfine structure which is not resolved except in a highly diluted salt.

PART I

1. INTRODUCTION

In the early applications of the method of paramagnetic resonance it is natural to choose substances whose spectra offer the greatest chance of a simple theoretical explanation. This requires salts whose crystallographic structure has been accurately determined; in addition, they must be fairly 'dilute' in the magnetic sense, in order to give narrow absorption lines, and to minimize the effects of exchange forces. In the iron group this suggests the use of the double sulphates: the alums, for the trivalent ions, and the monoclinic Tutton salts for the divalent ions. The chemical formulae are of the type $M'M(SO_4)_2 \cdot 12H_2O$ and $M''M_2(SO_4)_2 \cdot 6H_2O$ respectively, where M' is the trivalent ion, M'' the divalent ion and M a monovalent diamagnetic ion. There exists, therefore, for each paramagnetic ion a series of salts formed by using different monovalent ions, principally K, NH_4 , Rb, Cs and Tl; further variety may be obtained by using selenates instead of sulphates.

The simplest paramagnetic ion is Cu^{++} , $3d^9$, in which the orbital momentum is effectively quenched and the spin $s = \frac{1}{2}$ corresponds to a single electron. One of its salts, $CuSO_4 \cdot 5H_2O$, has been investigated by Baggeley & Griffiths (1948), who find that exchange forces play a dominant role in determining the behaviour of the energy levels. It is therefore particularly interesting to examine the more dilute Tutton salts, in which the crystalline electric field should be remarkably similar to that in the simple sulphate. This paper reports the results of measurements on seven copper Tutton salts, five sulphates and two selenates.

2. CRYSTALLOGRAPHY AND SUSCEPTIBILITY

The Tutton salts form a monoclinic series, in which the three crystallographic axes (a , b , c) are very closely in the ratio (3, 4, 2). The b -axis is normal to the plane

* Dr Penrose died suddenly at Leiden on 28 April 1949 while this paper was in the press.

containing a and c ; the a -axis makes an angle of 105° with the c -axis, measured positively in the anti-clockwise direction from the latter (see figure 1). No X-ray measurements on the copper salts exist, but following Polder (1942) we shall assume that the structure is similar to that of $\text{Mg}(\text{NH}_4)_2(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$, investigated by Hofmann (1931). The unit cell contains two molecules, one derived from the other by a translation from the point $(0, 0, 0)$ to $(\frac{1}{2}, \frac{1}{2}, 0)$ followed by a reflexion in the ac -plane. The divalent metallic ions lie in the points mentioned, surrounded by an octahedron of water molecules, of which four lie very nearly in a square each at a distance of 1.9 \AA from the Mg^{++} , and the other two at 2.15 \AA . The crystalline electric field should therefore have approximately tetragonal symmetry about an axis along the line joining the more distant waters. The two tetragonal axes in unit cell are equally inclined to the ac -plane, making an angle (α) of about 25° with it in the Mg^{++} salt. This angle is probably somewhat different in the Cu^{++} salts owing to the increased size of the ion.

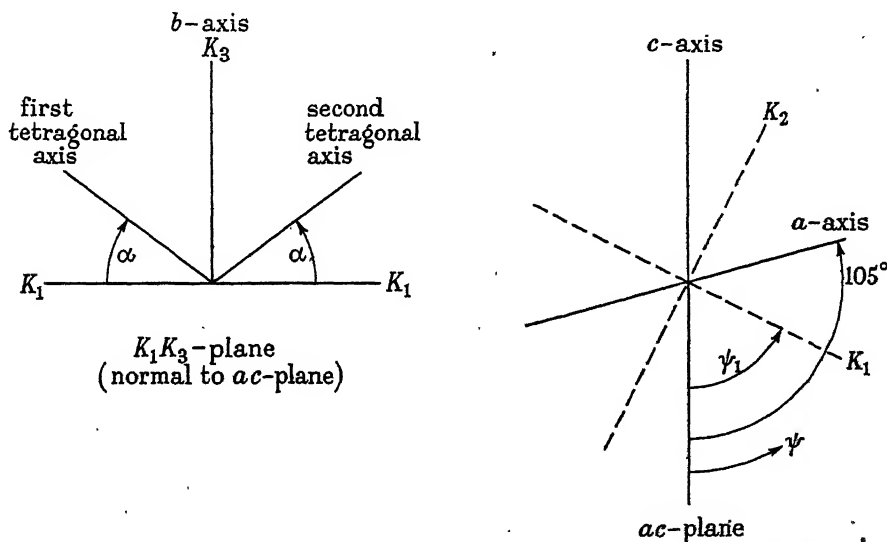


FIGURE 1. Crystallographic (a , b , c) and magnetic (K_1 , K_2 , K_3) axes.

On these assumptions, Polder shows that the orbital levels of the Cu^{++} ion should split into three singlet levels F_1 , F_3 and F_4 ; and one level, F_5 , doubly degenerate except for a small splitting due to the spin-orbit coupling. F_3 lies lowest, some $10,000 \text{ cm.}^{-1}$ below the next level.

Introduction of the spin of $\frac{1}{2}$ makes F_3 doubly degenerate, with a magnetic moment $g\beta\sqrt{(\frac{1}{2} \cdot \frac{3}{2})}$. Owing to the residual spin-orbit coupling, the value of g is not exactly 2, and is anisotropic, with the values

$$\begin{aligned} \text{parallel to the tetragonal axis:} \quad g_{\parallel} &= 2 \left(1 - \frac{4\lambda}{F_4 - F_3} \right), \\ \text{perpendicular to the tetragonal axis:} \quad g_{\perp} &= 2 \left(1 - \frac{\lambda}{F_5 - F_3} \right), \end{aligned} \quad (1)$$

where λ is the spin-orbit coupling coefficient $= -852 \text{ cm.}^{-1}$. Since F_5 lies above F_4 , g_{\parallel} will be greater than g_{\perp} , and the corresponding susceptibilities per g. ion along and normal to the tetragonal axis will be

$$\left. \begin{aligned} \chi_{\parallel} &= \frac{0.376}{T} \left(1 - \frac{4\lambda}{F_4 - F_3} \right)^2 + \frac{2.1}{F_4 - F_3}, \\ \chi_{\perp} &= \frac{0.376}{T} \left(1 - \frac{\lambda}{F_5 - F_3} \right)^2 + \frac{0.53}{F_5 - F_3} \end{aligned} \right\} \quad (2)$$

where F_3 , F_4 and F_5 are in cm.^{-1} .

Combining these results for the two ions in unit cell, one finds that of the principal axes of the ellipsoid of susceptibility, two lie in the ac -plane, one (K_1) being the bisector of the tetragonal axes, the other (K_2) being perpendicular to both these axes, while the third (K_3) lies along the crystallographic b -axis, which forms the other bisector of the tetragonal axes (figure 1). The corresponding susceptibilities are

$$\left. \begin{aligned} \chi_1 &= \chi_{\parallel} \cos^2 \alpha + \chi_{\perp} \sin^2 \alpha, \\ \chi_2 &= \chi_{\perp}, \\ \chi_3 &= \chi_{\parallel} \sin^2 \alpha + \chi_{\perp} \cos^2 \alpha, \end{aligned} \right\} \quad (3)$$

where α is the angle between either tetragonal axis and the ac -plane. From the susceptibility measurements of Hupse (1942) on $\text{CuK}_2(\text{SO}_4)_2 \cdot 6\text{H}_2\text{O}$ over the range 290 to 1.5°K , Polder deduces the following values:

$$g_{\parallel} = 2.44, \quad g_{\perp} = 2.05, \quad \alpha = 40^\circ,$$

which are in reasonable agreement with his calculations on the crystalline electric field. Measurements over a more restricted temperature range by Krishnan, Chakravorty & Banerjee (1933) Krishnan & Mookherji (1938) on other copper Tutton salts suggest that their behaviour is similar to that of the potassium salt.

3. THE PARAMAGNETIC RESONANCE EXPERIMENTS

In measurements of susceptibility only the net effect of the two ions taken together can be observed, and Polder points out that this leads to an ambiguity in the values of g_{\parallel} , g_{\perp} and α . *A priori*, it would have been equally valid to assume that the K_1 axis was perpendicular to the tetragonal axes, and that K_2 was a bisector. This would give $g_{\parallel} = 1.98$, $g_{\perp} = 2.27$ and $\alpha = 28^\circ$, and although the value of α agrees more closely with that deduced from the X-ray measurements of the magnesium salt, the agreement with the tetragonal field theory for the g -values would have been completely upset. Paramagnetic resonance experiments have the advantage that each ion in unit cell gives its own absorption line whose position is determined by the angle at which the applied magnetic field is inclined to the tetragonal axis. If this angle is β , then the effective g -value is given by

$$g^2 = g_{\parallel}^2 \cos^2 \beta + g_{\perp}^2 \sin^2 \beta \quad (4)$$

(this formula was kindly derived for us by K. W. H. Stevens). At a frequency $\bar{\nu} (\text{cm.}^{-1})$ the absorption line will occur at a field H_0 , where

$$hc\bar{\nu} = g\beta H_0 = g \frac{e\hbar}{4\pi mc} H_0.$$

If H_0 is measured in kilogauss, this reduces to

$$g = 21.4\bar{v}/H_0. \quad (5)$$

Thus the absorption lines for the different ions will occur (in general) in different fields; hence it is possible to measure g for each ion separately, and to locate its tetragonal axis, along which g should be a maximum. The values of g_{\parallel} and g_{\perp} can then be measured directly. In addition, by making measurements in the plane normal to a tetragonal axis it should be possible to determine how far the assumption of tetragonal symmetry is valid.

In this investigation the following measurements were carried out:

- A. The variation of g in the ac -plane.
- B. The value of g_3 along the b -axis (K_3).

The two tetragonal axes are equally inclined to the ac -plane, and to the b -axis; the absorption lines for the two ions will therefore coincide.* The maximum and minimum values of g (g_1 and g_2) in A determine the positions of K_1 and K_2 .

These two sets of measurements (A and B) are equivalent to the susceptibility measurements of Hupse and of Krishnan *et al.*, but they have the advantage that there is no correction for diamagnetism or temperature-independent paramagnetism. These measurements provide no further check on Polder's theory, nor do they identify the plane containing the tetragonal axes. The necessary evidence is furnished by the following experiments, which illustrate the particular advantages of the paramagnetic resonance method.

- C. The variation of g in the planes (K_1K_3) and (K_2K_3).
- D. The variation of g in a plane normal to a tetragonal axis.

If Polder's theory is correct, the tetragonal axes lie in the plane (K_1K_3), and two resolved lines should appear in this plane. On the other hand, only one line should appear in the (K_2K_3) plane, since it is equally inclined to both tetragonal axes. From the values g_1, g_2, g_3 for the principal susceptibility axes, the values of g_{\parallel}, g_{\perp} and α can be calculated. These can be compared with those directly measured.

Measurements A and B were made most conveniently at a wave-length of 3.2 cm. as also were the initial measurements of C. These were sufficient to show that two lines were obtained in the (K_1K_3) plane and only one in (K_2K_3). The two lines were not well resolved, and to make accurate measurements, a wave-length of 1.3 cm. was used for C and D.

These four sets of measurements provide a complete check of Polder's theory, and the results are presented in § 5 of this paper. In addition to the positions of the resonance lines, which determine the values of g , there is another parameter whose study is of great interest. This is the line width and line shape, which the preliminary experiments showed to vary markedly with orientation of the crystal and from salt to salt. The discussion of this question is given together with the results of the measurements, in part II.

* Since the one ion in unit cell is derived from the other by a reflexion in the ac -plane (plus a translation), the two absorption lines in this plane will coincide whatever the symmetry of the crystalline field.

4. METHOD AND APPARATUS

In view of the low susceptibility of the copper Tutton salts, and the fact that only small single crystals could readily be grown, it was decided to make the measurements principally at 90° K. At this temperature the dielectric losses are insignificant, and it is not necessary to grind the crystals down to reduce the amount protruding into the r.f. electric field, as is generally the case at room temperature.

The apparatus used is simple, and merits only a brief description. That used at 3 cm. wave-length has already been outlined (Bleaney & Penrose 1948); the apparatus for 1.3 cm. is essentially the same. A fixed-tuned cavity, one-half a wave-length long and resonant in the H_1 mode, is excited through hole-couplings from rectangular wave-guides constructed of thin-walled german silver tubing. It is cooled by means of liquid oxygen contained in a small dewar with a narrow tail. By this means a gap of 24 mm. between pole-faces tapered to $1\frac{1}{2}$ in. could be used in the magnet, giving fields up to 13 kG.

The experimental method is as follows at both wave-lengths. Power is fed to the cavity from a reflex klystron oscillator whose frequency is adjusted to the cavity resonance. A small fraction of the power entering the cavity is fed out through a second coupling to a silicon-tungsten rectifier crystal, whose d.c. current is read by a sensitive galvanometer. This current is closely proportional to the r.f. power incident on the rectifier, and hence to the square of the magnification factor Q of the cavity. When the latter is lowered because of absorption of power by the paramagnetic crystal in the cavity, the rectified current falls from its initial value δ_0 to some value δ_1 , and the paramagnetic absorption coefficient γ (in arbitrary units) can be calculated from the formula

$$\gamma = \sqrt{\left(\frac{\delta_0}{\delta_1}\right)} - 1.$$

To reduce γ to absolute units is difficult, requiring a measurement of Q and the filling factor of the crystal in the resonator; fortunately, the main interest lies in the shape rather than the absolute intensity of the absorption curves, and this reduction was not carried out.

In an experiment, a variable d.c. magnetic field H_0 is applied normal to the r.f. magnetic field, and the galvanometer reading δ is observed as a function of H_0 . At each reading the frequency of the klystron is adjusted to be in exact resonance with the cavity. This is necessary in order to correct for slight frequency fluctuations in the oscillator, and for the detuning of the cavity owing to anomalous dispersion in the paramagnetic salt. The background of fluctuation corresponded to a few per cent of the maximum of an absorption line, the shape of which could therefore be determined well out into the wings.

The magnetic field was calibrated initially by means of a search coil and flux-meter, and then over the limited ranges involved in the measurement of g by means of a ballistic galvanometer, which was standardized using the reversal of a known current in an accurate mutual inductance. The error in H_0 is thought to be less than 1 %.

5. RESULTS—VARIATION OF g

The first measurement (A) was of the variation of g in the ac -plane, at intervals of 20° . From equation (4) it follows that, if the magnetic field is applied at an angle ψ to the c -axis, and the K_1 axis makes an angle ψ_1 with this axis, then

$$g^2 = g_1^2 \cos^2(\psi - \psi_1) + g_2^2 \sin^2(\psi - \psi_1), \quad (4a)$$

where g_1 and g_2 are the values of g along K_1 and K_2 respectively. In all cases it was found that the points could be fitted within experimental error by a curve of the type (4a). A typical curve is shown in figure 2 for copper ammonium selenate. It should be noted that in these experiments ψ is measured in an anti-clockwise sense from the c -axis, as in figure 1, the a -axis lying at an obtuse angle $\psi = 105^\circ$. This corresponds to the notation used by Krishnan *et al.* (1933) and later Indian workers, and is of opposite sign to that of earlier measurements of Jackson (1923, 1926), Rabi (1927) and Bartlett (1932).*

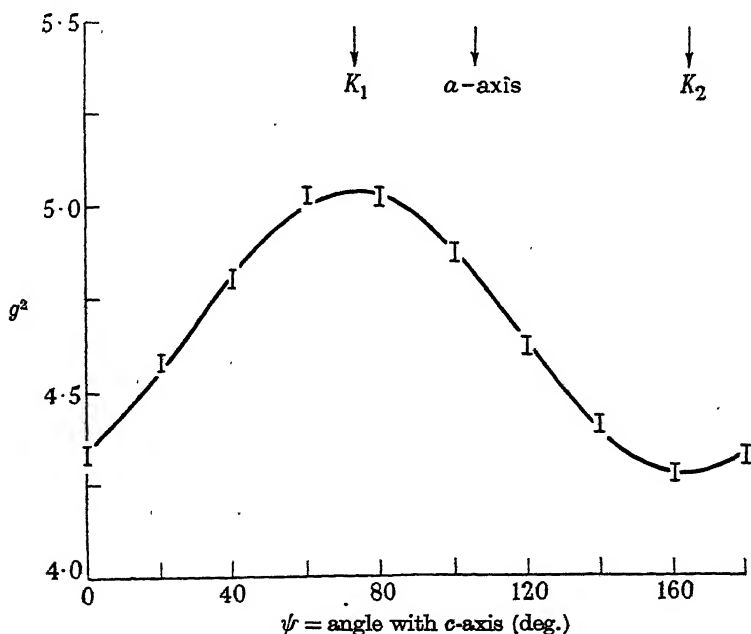


FIGURE 2. Variation of g^2 in the ac -plane for $\text{Cu}(\text{NH}_4)_2(\text{SeO}_4)_2 \cdot 6\text{H}_2\text{O}$.

The results of the measurements (A) and (B) on the seven salts examined are shown in table 1. Column 2 shows the values of ψ deduced from the positions of the

* The comparison of these measurements reveals some considerable confusion. Bartlett (1932) points out that in Rabi's (1927) paper the values of θ (our ψ_1) are interchanged for the potassium and ammonium salts of both nickel and cobalt. From a later correction (Bartlett 1933) it appears that this is true also of the copper salts; at the same time Bartlett accepts Krishnan's correction that his own values of θ should be reversed in sign. This makes them in good numerical agreement with Krishnan's own results, with the same sign, though they appear to have been measured in the opposite sense. In fact, the values obtained by both Rabi and Bartlett suggest that they were really measured in the same sense as Krishnan's; they would then agree well also with the results reported in this paper.

maximum and minimum values of g in the ac -plane; the accuracy is probably not greater than $(\pm)3^\circ$. For comparison the measurements of Krishnan *et al.* (1933) for the first two salts, and of Krishnan & Mookherji (1938) for the others are quoted in column 3.

TABLE 1

salt	ψ_1		g_1	$g_2 = g_\perp$	g_3	g_1	α	
	(this paper)	(Krishnan)					(this paper)	(Krishnan)
sulphates:								
potassium	105	102	2.26	2.05	2.22 _s	2.43	42	42
ammonium	65	77	2.27	2.06	2.20	2.40	39	42
rubidium	105	109	2.29 _s	2.07	2.23	2.45	40	40
thallium	112	108	2.26	2.06	2.20	2.39	39½	41
caesium	114	108	2.28	2.06	2.22	2.43	40	39
selenates:								
ammonium	72	67	2.24	2.06 _s	2.17	2.34	37½	39
potassium	73	126	2.26 _s	2.04 _s	2.17	2.38	37	41

A rough check was obtained on the values of ψ_1 by suspending the crystal by a torsionless fibre in a magnetic field, so that it was free to rotate about the b -axis. The K_1 axis then sets parallel to the field, and its position can be observed directly. The values of ψ_1 , obtained thus at room temperature, agreed closely with those given in column 1, except for the ammonium sulphate, for which the value was 77° . This latter value agrees with Krishnan *et al.* (1933); since for this salt a variation of ψ_1 with temperature has been reported (Bartlett 1932), a crude attempt was made to observe ψ_1 for this salt at 90° K. The value obtained was 69° , and a definite rotation could be observed as the crystal warmed up. The reason for the very wide discrepancy in the values of ψ_1 for the potassium selenate is unknown.

The next three columns give the values of g_1 and g_2 for the ac -plane, together with the value of g_3 obtained by measurement with the magnetic field directed along the b -axis. From these one can obtain the values of g_1 and α , assuming Polder's theory to be correct. From equation (4) it can easily be shown that

$$\left. \begin{aligned} g_1^2 + g_3^2 &= g_{||}^2 + g_\perp^2, \\ \cos 2\alpha &= (g_1^2 - g_3^2)/(g_{||}^2 - g_\perp^2), \\ g_\perp &= g_2. \end{aligned} \right\} \quad (6)$$

with, of course,

It will be seen that the values of ψ_1 tend to cluster around either 110 or 70° , and the values of α are slightly higher for the former group than for the latter. There is no similar correlation in the case of g_1 and g_\perp , which are, of course, sensitive to the magnitude of the crystalline electric field. The estimated accuracy of the values of g is $\frac{1}{2}$ to 1% and in α about 1° . The latter depends only on relative values of g , which are probably rather better than the absolute values, which depend on the accuracy of the calibration of the ballistic galvanometer used to measure the magnetic field.

Since in deducing the values of g_1 , g_\perp and α the same assumptions have been made as by Polder and by Krishnan, it is convenient to compare the results with those of other methods now, before reporting the other experimental results. Only for copper

potassium sulphate has the susceptibility been measured over a sufficient temperature range to obtain accurate values of the Curie constants (Hupse 1942). From these one finds

$$g_{\parallel} = 2.44, \quad g_{\perp} = 2.05 \quad \text{and} \quad \alpha = 40^{\circ},$$

which are in very good agreement with those reported in table 1.

It is possible to obtain values of α for all the salts from the anisotropy measurements at room temperature of Krishnan *et al.* (1933, 1938), if the assumption is made that the diamagnetic susceptibility is isotropic. The theory of Polder shows that the temperature-independent paramagnetism has tetragonal symmetry about the same axis as the temperature-dependent portion. Consequently, one can calculate α from the total anisotropy at any temperature, since from equation (3) it follows that

$$\cos 2\alpha = \frac{(\chi_1 - \chi_3)}{2(\chi_1 - \chi_2) - (\chi_1 - \chi_3)}.$$

The values of α obtained by this means from Krishnan's measurements are given in the last column of table 1 (see also Mookherji 1945). The agreement with our values is close, though not exact. It is, however, difficult to estimate the error that may arise from the assumption of diamagnetic isotropy in reducing Krishnan's values; and small changes in α with temperature may occur.

Experiments C and D. The variation of g in the planes (K_1K_3) and (K_2K_3) was not studied systematically, but by making experiments of an angle of about 45° to K_3 in the two planes at a wave-length of 3 cm., it was easy to verify that the two tetragonal axes lie in the plane (K_1K_3) . The absorption curve shows two peaks in this plane, but only one in (K_2K_3) . To obtain better resolution of the two peaks, observations were made at a wave-length of 1.3 cm. The separation between the two lines increases linearly with frequency, whereas the width remains substantially constant. It is thus possible to obtain a direct measurement of g_{\parallel} and also of g_{\perp} in the plane of the tetragonal axes instead of along K_2 . A typical curve, for copper ammonium sulphate, is shown in figure 3. The magnetic field is directed along one tetragonal axis, giving the low-field peak, and at an angle of just under 80° to the other axis. By a small rotation of the crystal the value of g_{\perp} can be determined. The results of measurements of this type are given in table 2.

For convenience, the values of g_{\parallel} and g_{\perp} ($= g_2$) from table 1 are repeated. It will be seen that the differences between the two sets of values are at first sight hardly outside the experimental error of $\pm 1\%$ except in the first two salts. In these there is a considerable discrepancy between the values of g_{\perp} in the (K_1K_3) plane and g_{\perp} ($= g_2$) in the (K_1K_1) plane, both of which are measured directly. That this is not due to errors in the magnetic field calibration or the measurement of wave-length was checked by measurements of g_2 at a wave-length of 1.3 cm., which gave values in excellent agreement with those obtained at 3.2 cm. It is apparent that the differences must be due to a departure from tetragonal symmetry; this is probably true of all the salts, but is more marked in the potassium and ammonium sulphates.

Since the two values of g_{\perp} obtained may not represent the extreme values when there is an appreciable departure from tetragonal symmetry, an investigation was made of the variation in a plane normal to one tetragonal axis for the potassium

sulphate. As two absorption lines of varying separation would be observed in this plane except along K_2 , it was essential to use a short wave-length, 1.3 cm., for this purpose. Measurements were made in four positions, along K_2 , and at angles of 45, 90 and 135° to it; the results are shown in table 3.

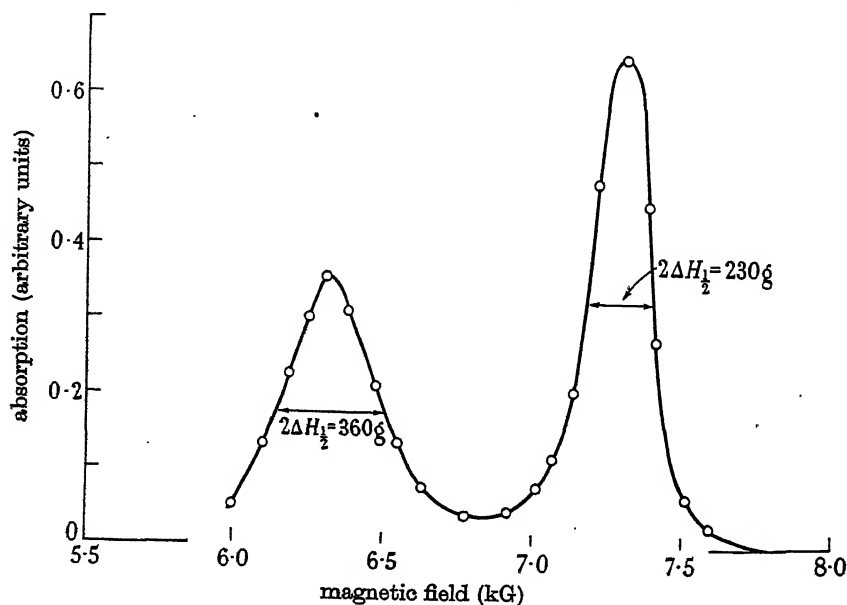


FIGURE 3. The resolved lines in the (K_1 , K_3) plane at 1.3 cm. wave-length for copper ammonium sulphate. Left-hand curve, tetragonal axis parallel to H ; right-hand curve, tetragonal axis at 80° to H .

TABLE 2

salt	in plane of tetragonal axes		from table 1	
	g_{\parallel}	g_{\perp}	g_1	g_2
sulphates:				
potassium	2.36	2.12	2.43	2.05
ammonium	2.45	2.12	2.39	2.06
rubidium	2.45	2.11	2.45	2.07
thallium	2.40	2.08	2.39	2.06
caesium	2.43	2.08	2.43	2.06
selenates:				
ammonium	2.39	2.07 _s	2.34	2.06 _s
potassium	2.38	2.07	2.38	2.04 _s

TABLE 3. VARIATION OF g IN PLANE NORMAL TO A TETRAGONAL AXIS FOR COPPER POTASSIUM SULPHATE

angle to K_2	observed values	average
0°	2.05, 2.05	2.05
45°	2.12, 2.13	2.12 _s
90°	2.10, 2.12, 2.12 _s	2.11 _s
135°	2.06 _s , 2.05 _s	2.06

The results show that the crystalline electric field must have an appreciable rhombic component, since the latter determines the departure of the g -value from

the free-spin value of 2, and there is a variation of at least a factor 2 in this departure. Consequently, it is not surprising that the values of g deduced on the assumption of tetragonal symmetry do not agree exactly with those measured directly. It follows also that the values of α may be slightly in error, though this does not affect the comparison with Krishnan's values, which rest also on the assumption of tetragonal symmetry. The slight misalignment of the magnetic field which would be caused by an error in α should not result in any appreciable error in the direct measurement of g_{\parallel} and g_{\perp} , since they are extreme values. An error in alignment of 7° would cause a change in g of only 0.01.

Although these measurements have shown that there is an appreciable departure from tetragonal symmetry in some of the salts, it is, nevertheless, convenient generally to ignore it because of the increased complication it introduces. For most purposes the error introduced will be small, since the value of g_{\parallel} is markedly greater than the value of g_{\perp} for all the salts. This arises partly because of the factor 4 which occurs in the expression for g_{\parallel} and not in that for g_{\perp} (equation (1)), and partly because the splitting which occurs in the denominator is smaller in the former expression. From the results reported in this paper it is possible to calculate these splittings for each salt. They will be substantially the same as those obtained by Polder from Hupse's measurements, and in view of the lack of detailed knowledge of the crystal-line field the variation for different salts would have only academic interest.

In view of the departure from tetragonal symmetry, it is at first sight surprising that the variation of the g -values in the ac -plane can be adequately represented by equation (4a). The values of g_1 , g_2 and ψ_1 are, however, adjusted to give the best fit, and the agreement would not be so good if the proper values deduced from g_{\parallel} and g_{\perp} were used. The expected discrepancy is revealed by measurements in the (K_2K_3) plane, where the constants are already fixed. When g is measured in this plane in directions at equal angles to the K_3 (b) axis, but on either side of it, the values obtained are unequal, though the angle with the tetragonal axes is the same in either case. This can only be explained by a rhombic component.

CONCLUSION

In the first approximation the variation of the g -values confirms Polder's theory, based on the assumption of two ions in unit cell each subjected to a crystalline electric field of tetragonal symmetry. The separate study of each ion made possible by the paramagnetic resonance method shows, however, that in some of the salts there is an appreciable departure from tetragonal symmetry. Thus in copper potassium sulphate the experimental results given in tables 2 and 3 lead to a rhombic anisotropy with the following principal values:

$$G_1 = 2.14, \quad G_2 = 2.04, \quad G_3 = 2.36.$$

The axis of G_3 (the 'tetragonal' axis) is at 41° to K_1 (in the plane of K_1 and K_3), while G_1 and G_2 make angles of approximately 70 and 160° to K_2 in the plane normal to G_3 . These values lead to

$$g_1 = 2.26, \quad g_2 = 2.05, \quad g_3 = 2.23,$$

in close agreement with the observed values (table 1) for the axes of principal magnetic susceptibility of the salt. From Hupse's measurements of the susceptibility the following values are obtained

$$g_1 = 2.27, \quad g_2 = 2.05, \quad g_3 = 2.21.$$

The agreement is within the experimental error.

A reference should be made here to the earlier theory of Jordahl (1934). The latter ignores the presence of two ions in unit cell and requires therefore a rhombic field to explain the different susceptibilities along the three axes K_1 , K_2 and K_3 . The constants of this rhombic field are chosen to give a small splitting (*c.* 300 cm.⁻¹) of the orbital doublet, in order to explain the fall of the Curie constants as the temperature is lowered, observed by Bartlett (1932). The latter has been confirmed by Janes (1935) over a wider temperature range (300 to 82° K). This variation with temperature does not require a small orbital splitting, however, since it can be explained by the temperature-independent paramagnetism which makes an important contribution to the susceptibility at the higher temperatures. The magnitude of the temperature-independent susceptibility for powders of copper ammonium and potassium sulphates deduced from Janes's results is $0.1_2 \times 10^{-4}$ (per g.-ion), while the theoretical values of Polder (equation (2) of this paper) give $0.1_5 \times 10^{-4}$.

PART II. LINE WIDTH AND SHAPE

7. INTRODUCTION

In a preliminary survey of the paramagnetic resonance method (Bagguley, Bleaney, Griffiths, Penrose & Plumptre 1948) it was pointed out that in general the width of resonance lines may be expected to vary with the temperature, the magnetic field and the orientation of the crystal in the field. The first of these factors enters through the spin-lattice interaction, whose characteristic relaxation time is generally strongly temperature-dependent. In the case of two of the copper Tutton salts, the experiments of Broer & Kemperman (1947) give a relaxation time of about 3×10^{-7} sec. at 90° K, which would contribute to the line width only 1 gauss or so. At lower temperatures the rapid increase in the relaxation time would make this contribution quite negligible. A direct comparison of line width at 90 and 20° K was made in a number of experiments, and no change of line width outside the experimental error of a few gauss was observed in any salt. It seems justifiable therefore to attribute the whole of the observed line width at these temperatures to the effects of spin-spin interaction between the various copper ions. The simplest type of interaction is associated with the local magnetic field of the neighbouring dipoles, and should give a line width independent of the magnitude of the external magnetic field. This latter was found to be the case for fields between 1 and 8 kG, measurements being made at 8.6, 3.2 and 1.3 cm. wave-length. The dipole-dipole interaction is strongly dependent on the direction of the external magnetic field, however, as in the corresponding nuclear resonance experiment (Purcell, Bloembergen & Pound 1946). It was therefore of considerable interest to study the change of line width with the orientation of the crystal.

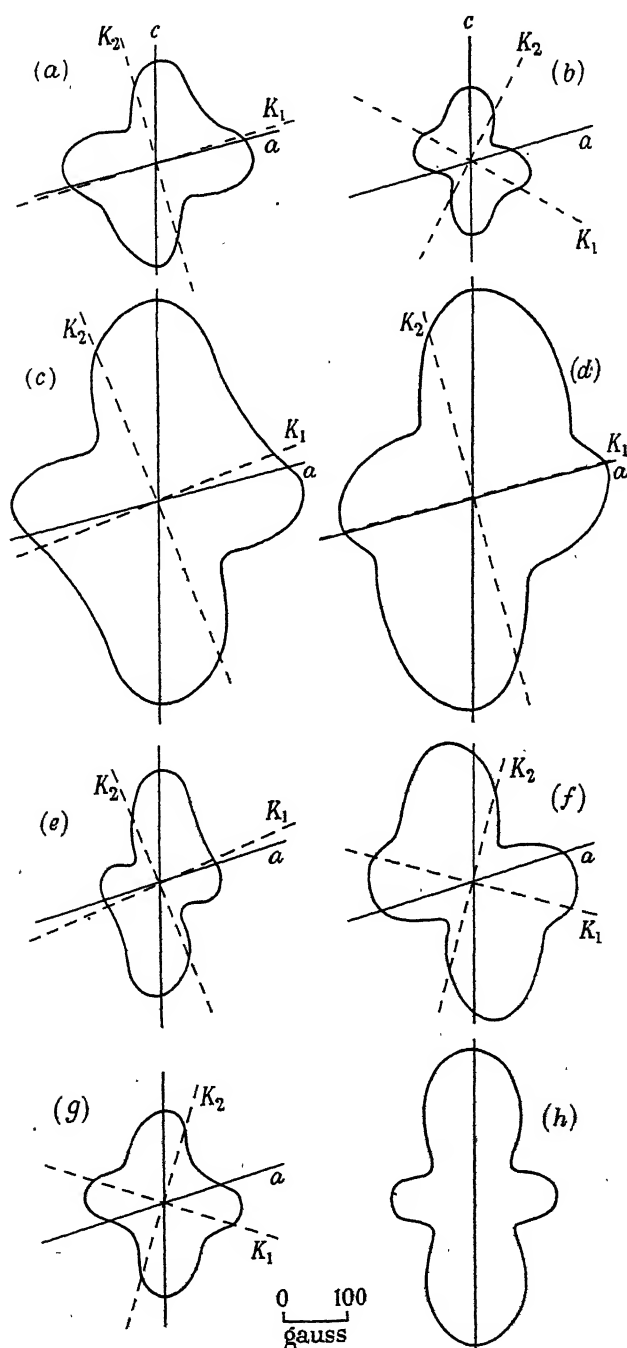


FIGURE 4. Line width (ΔH_2) in the ac -plane. (a), copper potassium sulphate; (b), copper ammonium sulphate; (c), copper caesium sulphate; (d), copper rubidium sulphate; (e), copper thallium sulphate; (f), copper ammonium selenate; (g), copper potassium selenate; (h), theoretical, assuming gaussian line shape.

For this purpose, measurement in the *ac*-plane suggests itself as most convenient and interesting. The nearest copper ions lie in this plane, and the two ions in unit cell give identical lines. Measurement of the line width was therefore combined with that of the *g*-value (experiments A), and was made at 3 cm. wave-length. The simplest quantity in measure is the half-width at half intensity ($\Delta H_{\frac{1}{2}}$), and the variation of this quantity with the direction of the external magnetic field in the *ac*-plane is shown in figure 4 in the form of a polar diagram, where the radius vector is proportional in length to ($\Delta H_{\frac{1}{2}}$). Though there is a striking variation in magnitude from salt to salt, the general form of the polar diagram is the same, having a main maximum along (or very close to) the *c*-axis, and a subsidiary maximum practically at right angles. The two minima are unequal, the deeper minimum lying sometimes on one side of the *c*-axis and sometimes on the other. Its position seems to be related to that of the principal axes of susceptibility (K_1, K_2), in that K_2 always lies between the deeper minimum and the *c*-axis. Interpretation of these results requires a detailed consideration of the interaction between the different copper ions.

8. MAGNETIC DIPOLE INTERACTION

The positions of the neighbouring copper ions are as follows, the distances given being those for the copper potassium sulphate. For the other salts the distances are slightly greater corresponding to the increase in the ionic volume, but in the same proportion, since the changes in the relative lengths and directions of the crystallographic axes are very small.

TABLE 4. POSITIONS OF NEAREST NEIGHBOURS

A. 2 ions at (0, 0, ± 1);	6.1 Å	E. 2 ions at (± 1 , 0, ± 1);	9.4 Å
B. 4 ions at ($\pm \frac{1}{2}$, $\pm \frac{1}{2}$, 0);	7.6 Å	F. 4 ions at ($\pm \frac{1}{2}$, $\pm \frac{1}{2}$, ∓ 1);	10.9 Å
C. 4 ions at ($\pm \frac{1}{2}$, $\pm \frac{1}{2}$, ± 1);	8.9 Å	G. 2 ions at (0, ± 1 , 0);	12.2 Å
D. 2 ions at (± 1 , 0, 0);	9.1 Å		

All other ions lie at distances of at least 12.2 Å.

From the simplest point of view the magnetic dipole interaction can be considered in the following way. The magnetic field of a neighbouring dipole μ gives rise to a component $\mu(1 - 3 \cos^2 \theta)/r^3$, where θ is the angle between the direction of the external field and the radius vector joining the two dipoles, which modifies the external field and so produces a spread in the Larmor precession frequency of the ions. Since the field falls off with the cube of the distance, the outstanding contribution will arise from the two ions A, which lie along the *c*-axis. The factor $(1 - 3 \cos^2 \theta)$ will therefore give a maximum along the *c*-axis, a subsidiary maximum at right angles half as great, and zeroes at angles of $\pm 55^\circ$ to the *c*-axis. This corresponds roughly to the observed polar diagrams, except that they are less anisotropic, and the zeroes are of course filled in by the fields of the more distant neighbours.

A closer approximation is obtained by taking the formula of van Vleck (1948) for identical spins:

$$(\Delta H^2)_{\text{av.}} = \frac{3}{2} g^2 \beta^2 S(S+1) \sum_j (1 - 3 \cos^2 \theta_{ij})^2 r_{ij}^{-6}, \quad (7)$$

where θ_{ij} = angle between magnetic field and line joining spins i and j , and $(\Delta H^2)_{av.}$ is the 'mean square moment'. In applying this to our case, we are ignoring the fact that, though the two ions in unit cell have the same g -factor in the ac -plane, their axes of precession are not quite parallel, being determined by the combined effect of the tetragonal field and the magnetic field. However, the major contribution comes from the ions at $(0, 0, \pm 1)$, which belong to the same sort as that at $(0, 0, 0)$ so that no serious error* is introduced by using the factor $\frac{3}{2}$ which appears in (7) for identical spins, in place of $\frac{1}{3}$ for non-identical spins. As a further simplification, we shall assume $g = 2.15$ for all directions in the ac -plane; this ignores a variation of $\pm 5\%$ which is determined by the position of the axes of susceptibility and is therefore different for each salt.

The contributions to the mean square moment from the ions listed in table 4 have been calculated at intervals of 15° in the ac -plane. An estimate of the contributions from more distant ions shows that they do not increase the mean square moment by more than a few per cent. To compare the calculated values of $\sqrt{(\Delta H^2)_{av.}}$ with the measured ΔH_i , some assumption must be made concerning the line shape. An approximation frequently used is that the shape is that of a gaussian distribution, $I = I_0 \exp[-\Delta H^2/2(\Delta H^2)_{av.}]$, for which $(\Delta H_i) = 1.18 \sqrt{(\Delta H^2)_{av.}}$. The curve h in figure 4 labelled 'theoretical' is obtained on this basis for copper potassium sulphate; it should be contracted slightly for the other salts in inverse ratio to their g -ionic volumes (listed in table 5). From the shape of the theoretical curve it will be seen that it is determined mainly by the two nearest neighbours at $(0, 0, \pm 1)$.

In some cases (for example, copper ammonium selenate) the line shape approximates very closely to that of a gaussian distribution. In other salts this is by no means the case, and two extreme shapes are shown in figure 5, the pointed line being obtained in copper ammonium sulphate. It is obvious that the small value of ΔH_i for the ammonium sulphate may be due to its pointed shape, and the large value for the caesium sulphate to its flattened shape. A numerical computation of $\Delta H_{av.}^2$ was therefore carried out for the absorption lines obtained in each salt along the c -axis; the results are shown in table 5 (column 4), together with the theoretical values (column 3). The g -ionic volumes are given in the second column, and the salts are arranged in order of increasing ionic volume. The experimental values† of ΔH_i are shown in the last column.

The results presented in table 5 indicate that the values of $\sqrt{(\Delta H^2)_{av.}}$ along the c -axis do not differ markedly from those calculated from equation (7), though the

* Professor M. H. L. Pryce has pointed out to us that the anisotropic g -values will produce a small modification of the factor $\frac{3}{2}$, since the contribution to the line width due to transitions caused by the rotating component of the field of the neighbouring ion involves a different g -value from that associated with the steady component. The error introduced is of the same order as the variation in the g -values.

† The probable errors in ΔH_i are $\pm 5G$; those given for $\sqrt{(\Delta H^2)_{av.}}$ are larger because they depend more on the shapes of the lines in the wings, where the intensity is low. In this respect the experimental values probably err on the low side rather than the reverse, since the tendency is to cut the tail off rather than to prolong it. The absorption is assumed to be zero when the magnetic field is far from resonance ($\pm 1 kG$), where the Q of the cavity becomes constant. Thus any tail of small intensity which extends so far is ignored, with corresponding reduction in the experimental values of $\Delta H_{av.}^2$.

values of ΔH_1 vary greatly. It is therefore interesting to examine whether a plot of the former quantity for the *ac*-plane would show a similar agreement. To obtain greater accuracy, a special set of measurements was carried out on copper ammonium selenate, for which the line shape is close to gaussian. The results are shown in figure 6, where the experimental points are drawn with a probable error of ± 5 G, and the

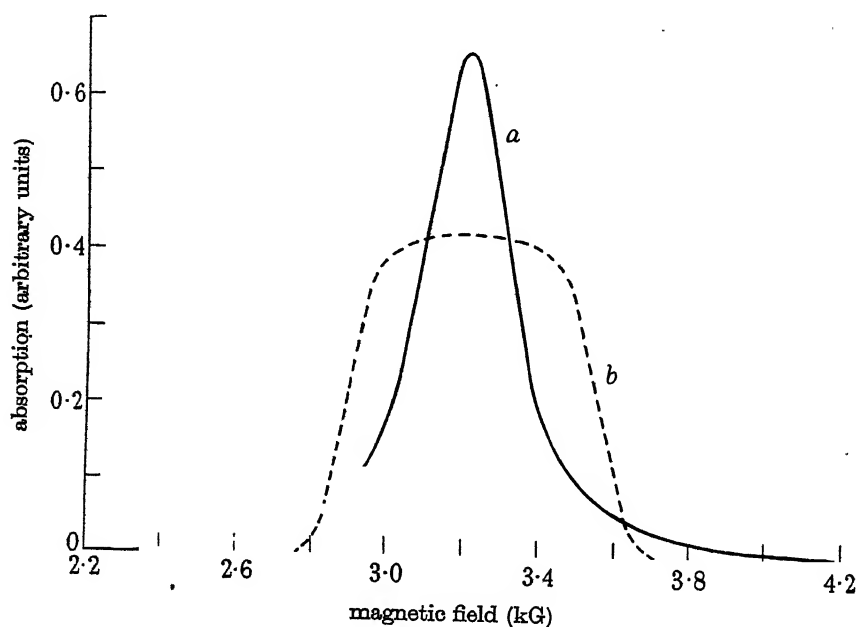


FIGURE 5. Line shape (along *c*-axis) for *a*, copper ammonium sulphate and *b*, copper caesium sulphate.

TABLE 5. LINE WIDTH ALONG THE *c*-AXIS

salt	g.-ionic volume (ml.)	$\sqrt{(\Delta H^2)_{av.}}$		$\Delta H_{\frac{1}{2}}$ exp. (gauss)
		theor. (gauss)	exp. (gauss)	
sulphates:				
potassium	197	196	214 ± 10	153
ammonium	206	187	226 ± 10	116
rubidium	212	181	235 ± 10	330
thallium	214	179	178 ± 10	172
caesium	219	175	198 ± 10	316
selenates:				
ammonium	224	171	180 ± 10	210
potassium	226	170	178 ± 10	141

broken line represents the theoretical value of the magnetic dipole interaction calculated by means of equation (7). Comparison with the curve of ΔH_1 (figure 4) for the same salt shows that the main features of the latter are reproduced in $\sqrt{(\Delta H^2)_{av.}}$ and they are not due merely to a change in line shape with direction. The discrepancies with the theoretical values are very considerable. In particular, the

sharp minimum at $+55^\circ$ to the c -axis is filled out, the experimental value being twice as great as the theoretical; and generally the width is greater than calculated, except near the K_2 axis. It is evident that other types of interaction between the magnetic ions of the same order as that due to the magnetic dipoles must be present in this salt.

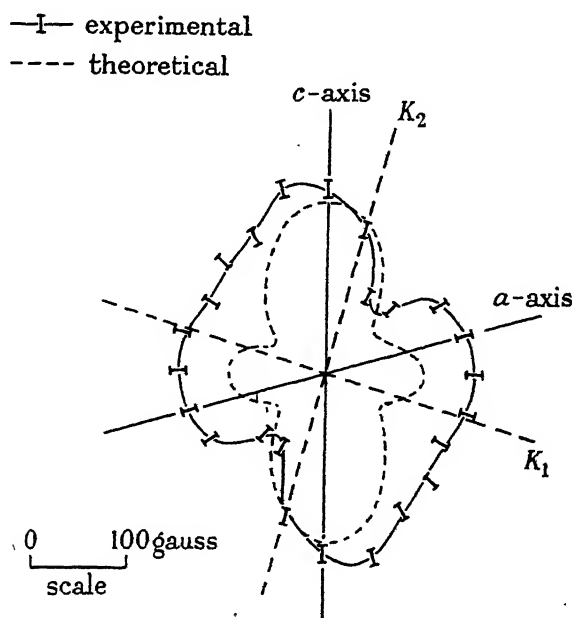


FIGURE 6. Polar diagram of $\sqrt{(\Delta H^2)_{av}}$ in the ac -plane for copper ammonium selenate.

9. ISOTROPIC EXCHANGE INTERACTION

In addition to the discrepancies revealed in figure 6, the magnetic dipole interaction offers no explanation of the widely divergent values of ΔH_i and the different line shapes, which seem to bear no relation to the ionic volumes. In no case do the lengths of the crystallographic axes depart by more than 3% from the simple (3, 4, 2) ratio, and the differences cannot be attributed to changes in the shape of the unit cell. It is necessary therefore to consider another type of interaction due to exchange forces. Although these are primarily connected with the orbital motion of the electrons, by virtue of the exclusion principle they manifest themselves as a coupling between the spin vectors proportional to the cosine of the angle between them. In the case of identical ions in a large magnetic field, this angle is independent of the direction of the magnetic field and the exchange forces are therefore isotropic. Gorter & van Vleck (1947) have pointed out that they give rise to a narrowing of the line in the centre and a broadening in the wings; that is, the line is more peaked and the value of ΔH_i is less than when magnetic dipole interaction alone is present. The exact line shape cannot be evaluated simply, but van Vleck (1948) has calculated the 'mean square moment' $(\Delta H^2)_{av}$ and the 'mean fourth moment' $(\Delta H^4)_{av}$. It turns out that the former is unaltered by exchange interaction; the discrepancy between

the experimental values of $(\Delta H^2)_{av.}$ and those calculated from magnetic interaction shown in figure 6 for copper ammonium selenate is therefore not explained by these isotropic exchange forces.

In view of the large changes in line shape between different salts it is interesting to consider what shape would be expected from magnetic interaction alone. The only parameter of line shape available for comparison with theory is the ratio (ρ) of the root mean fourth width $[(\Delta H^4)_{av.}]^{\frac{1}{4}}$ to the root mean square width $[(\Delta H^2)_{av.}]^{\frac{1}{2}}$. For a gaussian distribution this ratio is $3^{\frac{1}{2}} = 1.32$. Van Vleck shows that the curve should be somewhat blunter, the ratio being 1.25 for a cubic lattice of spins $S = \frac{1}{2}$. For the Tutton salts, as a first approximation one may assume that only the two nearest neighbours A at $(0, 0, \pm 1)$ make a significant contribution to the line width. Equation (24) of van Vleck (1948) then yields 1.18 for the ratio of the root mean fourth and second moments. It is interesting that the experimental value of this ratio for the curve for copper caesium sulphate shown in figure 5 is 1.1. This curve approximates in shape therefore to that which would be expected if magnetic dipole interaction alone were present,* though the actual value of the root mean square moment (table 5) is somewhat too large. For copper ammonium selenate the experimental value of ρ for the twelve absorption lines in the ac -plane used to obtain the root mean square width shown in figure 6 varies between 1.3₂ and 1.4₄. Thus in their case the line shape never departs greatly from the gaussian, and the polar diagram of the root mean square width (figure 6) has the same outline as the half-width (figure 4f). It is true of all the salts that no great change in line shape is discernible as the applied magnetic field is rotated in the ac -plane, and it is probable therefore that polar diagrams of $\sqrt{(\Delta H^2)_{av.}}$ would resemble those of ΔH_i for all the cases shown in figure 4, with, of course, a scaling factor dependent on the line shape.

Since the line shape is more or less isotropic, it may be attributed to simple exchange interaction. The great variation of line shape between different salts is not surprising, since under the influence of the crystalline electric field the orbital wave functions are strongly directional. The wide variation in the directions of the tetragonal axes from salt to salt indicates that a similar change in the overlapping of the orbital wave functions of neighbours, and hence of the magnitude of the exchange forces, may be expected. There is, however, no obvious correlation between line shape and the angle α which determines the position of the 'tetragonal' axis. In view of the presence of rhombic components of the crystalline field, this is not to be expected, especially as intervening atoms other than the water molecules (which mainly determine the crystalline field) will also exercise considerable influence on the orbital wave functions.

10. EXCHANGE INTERACTION BETWEEN THE DISSIMILAR IONS

Since the unit cell of the Tutton salts contains two copper ions subject to differently oriented crystalline fields, the spin vectors of these two ions will not precess about parallel axes when a magnetic field is applied, and the angle between the vectors will

* The rectangular shape may, however, be associated with the unresolved hyperfine structure (see §11).

vary with the direction of the external field. The effects produced then by exchange forces are illustrated by the (magnetically) very similar salt $\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$, in which a remarkable anomaly has been discovered by Bagguley & Griffiths (1948). When a magnetic field is applied in a direction making unequal angles with the two tetragonal axes, as in experiment C of part I of this paper, the resonance lines due to the two different ions are not resolved until magnetic fields of the order of 12 kG are used, although the line widths are less than a hundred gauss. This effect is attributed to exchange forces between the dissimilar ions (Pryce 1948) which are strong enough to overcome the separation in energy which the ions should possess in smaller magnetic fields. In addition, there is a broadening effect even when the two lines coincide, unless the magnetic field is applied in a direction perpendicular to both tetragonal axes, when even the dissimilar ions precess about identical axes.

It is obvious that in the Tutton salts such exchange forces, if present, are much weaker than in the single sulphate. Figure 3 shows how the lines are completely separated in copper ammonium sulphate in fields of 6 to 7 kG, and even at 3 cm. wave-length (fields of about 3 kG) the two peaks are resolved in all the seven Tutton salts investigated (see, for example, figure 7). It may be, however, that the exchange force between dissimilar ions does still produce a broadening of the lines, and in this connexion it may be significant that the discrepancy between the experimental root mean square widths and those calculated from magnetic dipole interaction revealed by figure 6 for copper ammonium selenate vanishes near the K_2 axis. This axis is perpendicular to both tetragonal axes, and exchange effects between dissimilar ions would not contribute to the line width. Whether they are large enough to explain the discrepancy when the magnetic field is applied in other directions cannot be estimated before further theoretical work has been carried out. They enhance only the contributions to $(\Delta H^2)_{\text{av}}$ from the dissimilar ions (B, C and F in table 4) which are rather small. In some directions they would have to be increased by a factor greater than 10 to account for the observed line widths in copper ammonium sulphate. It would be rather surprising if the effect were so large, since it arises only through the precession about slightly inclined axes.

11. EFFECT OF UNRESOLVED HYPERFINE STRUCTURE

Apart from these discrepancies it seems that none of the interactions considered hitherto can explain the unequal widths of the resolved absorption lines shown in figure 3. These lines are obtained by applying the magnetic field in a direction parallel to the tetragonal axis of one ion, when it will be an angle of about 80° to the tetragonal axis of the other ion. The latter gives the narrower line at the higher field (lower g -value). The spatial distribution of copper ions round each of these is precisely the same except that wherever X has a neighbour of the type X , Y has a neighbour of the type Y and vice versa. In this type of experiment, then, where the external magnetic field is in the same direction relative to the crystallographic axes for each ion, one would expect that the interactions with the neighbours should be identical, except that there may be a variation in the magnetic interaction associated with the different g -values of the two types of ion. The maximum difference this can

cause is of the same order as the difference in the g values ($\approx 2.4/2.05$), and its sign will depend on whether the major contribution to the width comes from similar or dissimilar ions. In this particular case, the contributions are roughly equal, and no significant difference in width would be expected. The experimental values of ΔH_i , however, are 184 G for the parallel ion, and 112 G for the ion whose tetragonal axis is nearly perpendicular to the magnetic field. Computation from the curve gives, 164 ± 20 and 114 ± 20 G respectively for $\sqrt{(\Delta H^2)_{av.}}$, while the theoretical value is approximately 90 G for each ion. Similar differences are found for the other salts a narrower absorption curve being obtained always for the perpendicular ion than for the parallel ion (see figure 7).

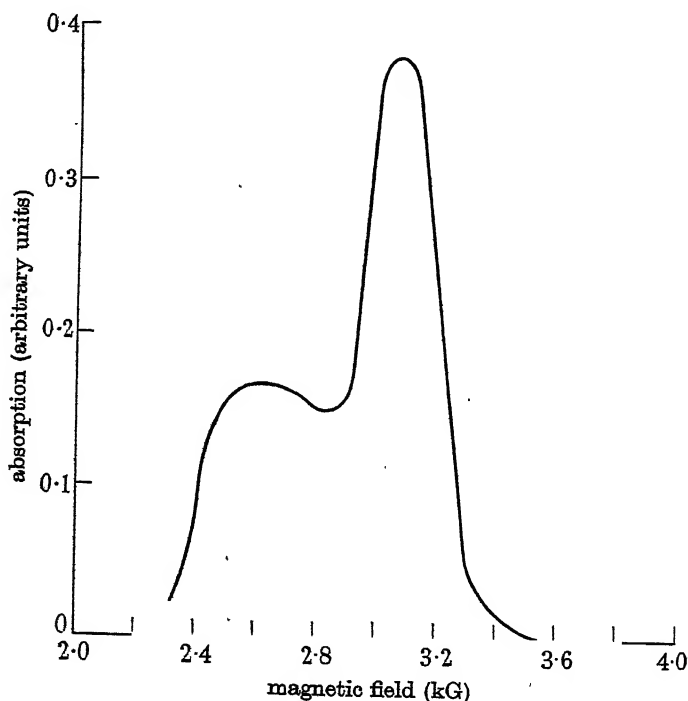


FIGURE 7. The resolved lines in (K_1, K_2) plane at 3.2 cm. wave-length from copper caesium sulphate. Left-hand curve, tetragonal axis parallel to H ; right-hand curve, tetragonal axis at 80° to H .

Since the only difference between the two ions is that their tetragonal axes are differently oriented with respect to the magnetic field, this difference in the mean square moments cannot be ascribed to magnetic dipole interaction or to isotropic exchange forces. The exchange forces between dissimilar ions may enhance the line width, but may be expected to affect either line equally, since physically they correspond to a rapid exchange of the two ions concerned. Thus there appears to be some effect which contributes to the mean square moments of the absorption lines, which is not explicable in terms either of magnetic interaction or exchange interaction between the spins. The magnitude of the contribution varies with the direction of the applied field, being greater where the g -values are higher. Further evidence

in support of this last statement is provided by measurements in the (K_2K_3) plane. At equal angles to, but on either side of the K_3 axis, the line widths are different. In copper ammonium sulphate at angles of $+30^\circ$ and -30° to K_3 , the values of ΔH_i are 87 and 59 G respectively. Since K_3 is identical with the crystallographic b -axis, contributions to the line width from magnetic dipole and exchange forces between similar or dissimilar ions should be identical in such directions equally inclined to it, as also should those from any effect with the tetragonal symmetry assumed for the crystalline field. The difference in the g -values of the two lines (2.19 and 2.13 respectively) shows, however, that there is an appreciable rhombic component of the crystalline field. The difference between the line widths is in the same direction as that in the g -values and could be explained by an effect with the same anisotropy as the crystalline field.

The nature of this effect has been revealed by subsequent work of Penrose at Leiden University. On using a crystal of copper ammonium sulphate greatly diluted with the corresponding magnesium salt, the single absorption line expected in the ac -plane breaks up into four equally spaced narrow lines, each with ΔH_i of about 15 G. The overall separation of the lines varies with the direction of the magnetic field, the maximum value of about 300 G being obtained in the K_1 direction, while near K_2 the lines are not completely resolved. This hyperfine structure is attributed to interaction between the electron spin and the copper nucleus, both of whose common isotopes have spin $\frac{3}{2}$ and nearly equal magnetic moments (2.226 and 2.385 nuclear magnetons (Pound 1948)). With the magnetic fields used, the spectrum corresponds to that of the Back-Goudsmit region in optical hyperfine spectra. The separation of the four components varies because the magnetic interaction between the electron and nuclear spins is averaged over the distribution of the electron cloud density in space, i.e. the orbital wave functions, which are determined by the crystalline electric field. If the latter has tetragonal symmetry, the separation should vary as $(1 + 3 \cos^2 \beta)^{\frac{1}{2}}$, where β is the angle which the magnetic field makes with the tetragonal axis. This is roughly consistent with the initial measurements of Penrose.

The separation of the hyperfine components is so large that they should be visible in the undiluted salts. The absorption curves, of which examples are given in figures 3, 5 and 7, show no trace of a fine structure. Its absence must be attributed to exchange forces, and perhaps also to the fluctuating components of the local magnetic field of neighbouring electron spins which shorten the lifetime of the electronic state sufficiently to 'average out' the interaction with the nuclear spin. The contribution of the latter to the mean square moment will remain, however, and it has been estimated by Pryce from the initial results of Penrose as

$$(\Delta H^2)_{av.} = 0.4 \times 10^4 (1 + 3 \cos^2 \beta) \text{ gauss}^2. \quad (8)$$

This formula can be applied to the two resolved lines (figures 2, 6), from which reasonable estimates of $(\Delta H^2)_{av.}$ can be obtained. The results are given in the table on p. 426.

The agreement is better than could be expected at this stage. Analysis of the difference between the mean square moments observed for copper ammonium

selenate in the *ac*-plane (figure 6) and those calculated from magnetic interaction shows that it also is of the same order as predicted by equation (8). The presence of an unresolved hyperfine structure offers therefore an adequate explanation of the deviation of the mean square moments from those expected from magnetic dipole interaction between the electron spins.

salt	angle with tetragonal axis	$(\Delta H^2)_{av. \text{ gauss}^2}$		
		experimental	magnetic interaction alone	magnetic interaction + hyperfine structure
copper ammonium sulphate	0°	$2.7 \pm 0.3 \times 10^4$	0.8×10^4	2.4×10^4
copper caesium sulphate	76°	$1.3 \pm 0.2 \times 10^4$	0.8×10^4	1.3×10^4
	0°	$2.3 \pm 0.2 \times 10^4$	0.7×10^4	2.3×10^4
	80°	$1.3 \pm 0.2 \times 10^4$	0.7×10^4	1.2×10^4

12. DISCUSSION

The assumption of exchange forces in the copper Tutton salts is not new, since they have been invoked by Opechowski (1948) to explain the anomalous specific heat of the potassium sulphate discovered by de Klerk (1946). The constant of the magnetic specific heat (CT^2/R) is 6.8×10^{-4} , which is some five times greater than the theoretical value for pure dipole interaction. Relaxation measurements by Bijl (1941) and Broer & Kemperman (1947) confirm this high value. De Klerk finds also that the susceptibility obeys a Curie-Weiss law with a Weiss constant of 0.052°K . Using this value, a tentative estimate by Opechowski suggests that at least ten neighbouring copper ions must co-operate in the exchange interaction. Reference to the list of neighbours, table 4 (Opechowski's list omits the ions labelled C and F in table 4) shows that this requires all the ions up to a distance of 9 \AA . The interaction is thus comparatively long range. Opechowski's theory neglects magnetic dipole interaction, however, which the results given in this paper show to be at least as important as exchange forces, as far as line width is concerned.

The hyperfine structure will also give rise to an additional specific heat, though insufficient to account for the observed value in copper potassium sulphate. Since the structure is 'averaged out' by the exchange interaction in the undiluted salt, its contribution to the specific heat may also be reduced. On the whole it appears that the abnormally high specific heat must be due in the main to the exchange interaction. This is supported by paramagnetic relaxation measurements in the liquid helium region by Benzie & Cooke (1949) who find the specific heat of copper rubidium sulphate, which gives the broadest absorption lines in our experiments, to be considerably smaller than that of the potassium or ammonium sulphates or the potassium selenate, all of which give fairly narrow lines.

In copper sulphate pentahydrate the specific heat at helium temperature has a marked anomaly at helium temperatures (Ashmead 1939) and the Weiss constant (0.7°K) is very large (Reekie 1939). The work of Bagguley & Griffiths shows that exchange forces are dominant as far as line width is concerned, the lines being narrower than in any of the Tutton salts, though the salt is magnetically more con-

centrated. The distance between nearest neighbours is not much less (5.5 \AA) than in the Tutton salts (6 \AA); it is therefore not surprising that exchange effects appear in the latter, though to a less degree, since there are only two ions at 6 \AA compared with eight in the single sulphate; in particular, the least distance between ions of dissimilar kind is 7.5 \AA in the Tutton salts, and there is no evidence of an anomalous absorption spectrum attributed to exchange between those ions, similar to that found by Bagguley & Griffiths.

13. CONCLUSIONS

(a) The change of line width with direction of the applied magnetic field is roughly consistent with that expected from magnetic dipole interaction, but the mean square moment is generally bigger.

(b) The line shape is more or less independent of the direction of the applied field in the *ac*-plane, but varies markedly from salt to salt. In a number of cases the lines are more peaked than can be explained by magnetic dipole interaction, an effect which may be attributed to isotropic exchange forces.

(c) The presence of exchange interactions is consistent with the abnormal specific heat of the spin-system, which is greater than can be accounted for by magnetic dipole interaction.

(d) The measured mean square moments show a departure from the calculated values which is greatest in directions where the *g*-values are greatest, i.e. the effect seems to have the same anisotropy as the crystalline electric field. It may partly be due to exchange interaction between dissimilar ions, but the presence of a hyperfine structure 'averaged out' by the exchange forces would give a contribution to the mean square moment of the required magnitude, and with the right anisotropy.

The authors wish to acknowledge their indebtedness to Professor M. H. L. Pryce and Mr K. W. H. Stevens for their contributions to the theoretical aspects; and to Mr D. M. S. Bagguley for help in growing crystals and for valuable discussion.

REFERENCES

- Ashmead, J. 1939 *Nature*, **143**, 853.
Bagguley, D. M. S., Bleaney, B., Griffiths, J. H. E., Penrose, R. P. & Plumpton, B. I. 1948 *Proc. Phys. Soc.* **61**, 542.
Bagguley, D. M. S. & Griffiths, J. H. E. 1948 *Nature*, **162**, 538.
Bartlett, B. W. 1932 *Phys. Rev.* **41**, 818.
Bartlett, B. W. 1933 *Phys. Rev.* **44**, 687.
Benzie, R. J. & Cooke, A. H. 1949 (Unpublished.)
Bijl, D. 1941 *Physica*, **8**, 461.
Bleaney, B. & Penrose, R. P. 1948 *Proc. Phys. Soc.* **60**, 395.
Broer, L. J. F. & Kemperman, J. 1947 *Physica*, **13**, 465.
de Klerk, D. 1946 *Physica*, **12**, 513.
Gorter, C. J. & van Vleck, J. H. 1947 *Phys. Rev.* **72**, 1128.
Hofmann, W. 1931 *Z. Kristallogr.* **78**, 279.
Hupse, J. C. 1942 *Physica*, **9**, 633.
Jackson, L. C. 1923 *Phil. Trans. A*, **224**, 1.
Jackson, L. C. 1926 *Phil. Trans. A*, **226**, 107.
Janes, R. B. 1935 *Phys. Rev.* **48**, 78.

- Jordahl, O. M. 1934 *Phys. Rev.* **45**, 87.
 Krishnan, K. S., Chakravorty, N. C. & Banerjee, S. 1933 *Phil. Trans. A*, **232**, 99.
 Krishnan, K. S. & Mookherji, A. 1938 *Phil. Trans. A*, **237**, 135.
 Mookherji, A. 1945 *Indian J. Phys.* **9**, 63.
 Opechowski, W. 1948 *Physica*, **14**, 237.
 Polder, D. 1942 *Physica*, **9**, 709.
 Pound, R. V. 1948 *Phys. Rev.* **73**, 523.
 Pryce, M. H. L. 1948 *Nature*, **162**, 539.
 Purcell, E. M., Bloembergen, N. & Pound, R. V. 1946 *Phys. Rev.* **70**, 988.
 Rabi, I. I. 1927 *Phys. Rev.* **29**, 174.
 Reekie, J. 1939 *Proc. Roy. Soc. A*, **173**, 367.
 van Vleck, J. H. 1948 *Phys. Rev.* **74**, 1168.
-

The theory of plane plastic strain for anisotropic metals

By R. HILL, *Cavendish Laboratory, University of Cambridge*

(Communicated by E. Orowan, F.R.S.—Received 4 February 1949)

A yield criterion and plastic stress-strain relations are formulated for anisotropic metals deformed under conditions of plane strain. The equations are shown to be hyperbolic, the characteristics coinciding with the directions of maximum shear strain-rate. When the anisotropy is uniformly distributed, the variation of the stresses along the characteristics is expressed in terms of elliptic functions, and geometrical properties of the field of characteristics are established. The theory is applied to the problem of indentation by a flat die.

1. INTRODUCTION

In an earlier paper (Hill 1948) a theory was proposed describing the macroscopic plastic behaviour of polycrystalline anisotropic metals. The theory was shown to be consistent with the experimental evidence then available, and has since been found to be in accord with more recent data (an account will be presented elsewhere). It seems worth while, therefore, to pursue the implications of the theory in greater detail.

The present paper is concerned with the two-dimensional problem of plane plastic strain. General methods of integrating the plane-strain equations for isotropic metals are well understood; the equations are hyperbolic, the characteristics being in the directions of maximum shear stress or shear strain-rate. Relations describing the variation of the stress and velocity components along the characteristics, together with various geometrical theorems, were established by Hencky (1923) and Geiringer (1930). The equations of plane strain for anisotropic metals are also hyperbolic, and the characteristics are in the directions of maximum shear strain-rate (Hill 1948). The main purpose of the present paper is to obtain the relations holding along the characteristics, and to examine whether there are simple geometrical properties analogous to the theorems of Hencky for isotropic metals.

2. THE EQUATIONS REFERRED TO THE AXES OF ANISOTROPY

The anisotropy is regarded as being due to a preferred orientation of the grain texture, and is assumed to have three mutually orthogonal planes of symmetry at every point. Their lines of intersection are called the axes of anisotropy. In the following analysis it is supposed further, for simplicity, that the anisotropy is uniformly distributed in magnitude and direction. This is, for example, approximately realized in a bar or strip cut from the central part of a cold-rolled sheet; the axes of anisotropy lie in the direction of rolling, transversely in the plane of the sheet, and normal to this plane.

Let (x, y, z) axes of reference be taken coincident with the axes of anisotropy. Let the state of plane strain be such that the z axis is normal to the planes of flow. Then, according to the theory formulated in the earlier paper (Hill 1948), the yield criterion takes the form

$$f(\sigma_{ij}) \equiv \left(\frac{FG + GH + HF}{F + G} \right) (\sigma_{xx} - \sigma_{yy})^2 + 2N\sigma_{xy}^2 = 1, \quad (1)$$

where F, G, H, N are parameters expressing the state of anisotropy. If X, Y, Z are the uniaxial tensile yield stresses in the directions of the axes of anisotropy, and T is the yield stress in shear in the (x, y) plane with respect to these axes, then

$$\left. \begin{aligned} 2F &= \frac{1}{Y^2} + \frac{1}{Z^2} - \frac{1}{X^2}, \\ 2G &= \frac{1}{Z^2} + \frac{1}{X^2} - \frac{1}{Y^2}, \\ 2H &= \frac{1}{X^2} + \frac{1}{Y^2} - \frac{1}{Z^2}, \\ 2N &= \frac{1}{T^2}. \end{aligned} \right\} \quad (2)$$

The stress σ_{zz} , normal to the planes of flow, is given by

$$\sigma_{zz} = (G\sigma_{xx} + F\sigma_{yy})/(G + F). \quad (3)$$

The stresses also satisfy the equations of equilibrium

$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xy}}{\partial y} = 0, \quad \frac{\partial \sigma_{xy}}{\partial x} + \frac{\partial \sigma_{yy}}{\partial y} = 0. \quad (4)$$

For reasons that will appear later it is convenient to write

$$c = 1 - \frac{N(F + G)}{2(FG + GH + HF)} \quad (-\infty < c < 1). \quad (5)$$

If N is greater than both $F + 2H$ and $G + 2H$, c is negative, while if N is less than both $F + 2H$ and $G + 2H$, c is positive. States of anisotropy for which N is intermediate to $F + 2H$ and $G + 2H$ appear not to be easily realizable in practice. c is zero if the material is isotropic, and also for states of anisotropy such that

$$N = F + 2H = G + 2H \quad (F \neq H).$$

With the notation of equation (5) the yield criterion can be rewritten as

$$f(\sigma_{ij}) \equiv \frac{(\sigma_{xx} - \sigma_{yy})^2}{4(1-c)} + \sigma_{xy}^2 = T^2. \quad (6)$$

The plane-strain tensile yield stress σ in the direction making an angle θ with the x axis is found by substituting

$$\sigma_{xx} = \sigma \cos^2 \theta, \quad \sigma_{yy} = \sigma \sin^2 \theta, \quad \sigma_{xy} = \sigma \sin \theta \cos \theta,$$

in (6), leading to
$$\sigma = 2T \left(\frac{1-c}{1-c \sin^2 2\theta} \right)^{\frac{1}{2}}. \quad (7)$$

It is evident that σ has equal values in any set of four directions $\pm \theta$, $\pm (\frac{1}{2}\pi - \theta)$, and hence that the angular variation of σ is symmetrical about the axes of anisotropy and about the 45° directions. The corresponding values (3) of σ_{zz} are, however, different unless $F = G$. If c is positive, σ has a minimum value $2T\sqrt{1-c}$ in the directions of the axes of anisotropy, and a maximum value $2T$ in the 45° directions; if c is negative, σ has a maximum value $2T\sqrt{1-c}$ in the directions of the axes of anisotropy, and a minimum value $2T$ in the 45° directions.

The tensor $\dot{\epsilon}_{ij}$, representing the rate of strain referred to the axes of anisotropy, is assumed to be related to the function $f(\sigma_{ij})$, governing yielding, by the equation

$$\dot{\epsilon}_{ij} = \lambda \frac{\partial f}{\partial \sigma_{ij}}, \quad (8)$$

where λ is a positive factor of proportionality which varies both in space and time. If other axes of reference are taken, the components of stress and strain-rate are related by a similar equation, the form of the function f being changed under the transformation because of the anisotropy. If v_x and v_y denote the components of velocity referred to the anisotropic directions, elimination of λ from (8) gives

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} = 0, \quad (1-c) \left(\frac{\partial v_x}{\partial x} - \frac{\partial v_y}{\partial y} \right) / \left(\frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x} \right) = \frac{\sigma_{xx} - \sigma_{yy}}{2\sigma_{xy}}. \quad (9)$$

The first equation expresses the incompressibility of the material, and the second is equivalent to a relation between the orientation ψ of a principal stress direction, with respect to the x axis, and the orientation ψ' of a principal strain-rate direction:

$$\tan 2\psi' = (1-c) \tan 2\psi. \quad (10)$$

Hence $\psi' = \psi$ only if $\psi = 0, 45$ or 90° ; this is a consequence of the fourfold symmetry of the angular variation of the tensile yield stress.

The five equations (4), (6) and (9) between the five unknowns σ_{xx} , σ_{yy} , σ_{xy} , v_x and v_y , involve only two parameters, namely, T , which is a measure of the average resistance to deformation, and c , which specifies the state of anisotropy in the planes of flow. T and c can be experimentally determined by two measurements, for example, in compression tests (under conditions of plane strain) at 0 and 45° to the axes of anisotropy. It is only necessary to know the separate magnitudes of the four parameters F , G , H , N (or X , Y , Z , T) if it is required to calculate σ_{zz} .

3. THE EXISTENCE OF CHARACTERISTICS

It will now be shown that there exist curves (characteristics) across which certain derivatives of the stress and velocity components may be discontinuous under suitable boundary conditions. Suppose that the stress distribution has been determined in some region to the left of a curve C (figure 1), and that it is required to find whether this solution can be continued to the right of C . Let Cartesian axes of reference (ξ, η) be taken coincident with the normal and tangent at some point P on C , and let ϕ_0 be the anti-clockwise rotation of the ξ axis from the x direction of anisotropy. The corresponding components of stress are denoted by $\sigma_{\xi\xi}$, $\sigma_{\eta\eta}$ and $\sigma_{\xi\eta}$.

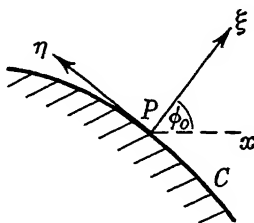


FIGURE 1. Co-ordinate axes for characteristic directions.

If the distribution of stress is assumed to be continuous, the derivatives $\partial\sigma_{\xi\xi}/\partial\eta$, $\partial\sigma_{\eta\eta}/\partial\eta$ and $\partial\sigma_{\xi\eta}/\partial\eta$ at the point P are continuous across C . Hence, from the equations of equilibrium, $\partial\sigma_{\xi\xi}/\partial\xi$ and $\partial\sigma_{\xi\eta}/\partial\xi$ are also continuous across C at P . Differentiation of the yield criterion $f(\sigma_{\xi\xi}, \sigma_{\eta\eta}, \sigma_{\xi\eta}) = 0$ with respect to ξ gives the equation

$$\frac{\partial f}{\partial \sigma_{\xi\xi}} \frac{\partial \sigma_{\xi\xi}}{\partial \xi} + \frac{\partial f}{\partial \sigma_{\eta\eta}} \frac{\partial \sigma_{\eta\eta}}{\partial \xi} + \frac{\partial f}{\partial \sigma_{\xi\eta}} \frac{\partial \sigma_{\xi\eta}}{\partial \xi} = 0, \quad (11)$$

from which to calculate $\partial\sigma_{\eta\eta}/\partial\xi$. The solution is unique, unless $\partial f/\partial\sigma_{\eta\eta} = 0$ at P , that is, unless $\dot{\epsilon}_{\eta\eta} = 0$, according to (8). In this case, further boundary conditions must be prescribed if the solution is to be continued across C ; these may be such that $\partial\sigma_{\eta\eta}/\partial\xi$ is discontinuous. C is therefore a characteristic if it coincides at every point with a direction of zero extension. Through each point pass two characteristics; the orthogonal directions of zero extension. Since the material is incompressible these are also the directions of maximum shear strain-rate (slip-lines), but not, in general, maximum shear-stress directions.

If, then, the tangent at P is a direction of zero extension, $\partial f/\partial\sigma_{\xi\xi} = -\partial f/\partial\sigma_{\eta\eta} = 0$. From (11), $\partial\sigma_{\xi\eta}/\partial\xi = 0$, and similarly, by differentiating the yield criterion with respect to η , it may be shown that $\partial\sigma_{\xi\eta}/\partial\eta = 0$. Hence, from the equations of equilibrium,

$$\frac{\partial \sigma_{\xi\xi}}{\partial \xi} = 0 = \frac{\partial \sigma_{\eta\eta}}{\partial \eta}, \quad (12)$$

when the (ξ, η) axes coincide with the directions of zero extension at P .

The existence of characteristics for the velocity components (v_ξ, v_η) may be similarly demonstrated. Supposing the velocity to be continuous, the derivatives $\partial v_\xi/\partial\eta$ and $\partial v_\eta/\partial\eta$ at P must be continuous across C . From the incompressibility

equation it follows that $\partial v_\xi / \partial \xi$ is also continuous. However, if $\partial f / \partial \sigma_{\xi\xi} = -\partial f / \partial \sigma_{\eta\eta} = 0$, there may be discontinuities in λ and $\partial v_\eta / \partial \xi$. The directions of zero extension, therefore, are also characteristics for the velocities. If the (ξ, η) axes coincide with the directions of zero extension at P ,

$$\frac{\partial v_\xi}{\partial \xi} = -\frac{\partial v_\eta}{\partial \eta} = 0. \quad (13)$$

The inclinations dy/dx of the characteristics relative to the x axis of anisotropy may easily be shown to be the roots of the equation

$$(\sigma_{xx} - \sigma_{yy})(dx^2 - dy^2) + 4(1-c)\sigma_{xy}dx dy = 0. \quad (14)$$

4. RELATIONS ALONG THE CHARACTERISTICS

For applications to special problems it is necessary to regard the characteristics as curvilinear axes, and to introduce the corresponding stress components $\sigma_{\alpha\alpha}$, $\sigma_{\beta\beta}$, $\sigma_{\alpha\beta}$, where the two families of characteristics are denoted by the symbols α and β . The families are distinguished by the convention that $\sigma_{\alpha\beta}$ shall be a positive quantity, to preserve the analogy with the isotropic theory. If ϕ is the anti-clockwise orientation of an α characteristic to the x axis of anisotropy, the yield criterion (6) becomes

$$f(\sigma_{\alpha\alpha}, \sigma_{\beta\beta}, \sigma_{\alpha\beta}) \equiv \frac{1}{(1-c)} [(\sigma_{\alpha\alpha} - \sigma_{\beta\beta}) \cos 2\phi - 2\sigma_{\alpha\beta} \sin 2\phi]^2 + [(\sigma_{\alpha\alpha} - \sigma_{\beta\beta}) \sin 2\phi + 2\sigma_{\alpha\beta} \cos 2\phi]^2 = 4T^2$$

on transforming the components of stress. The condition for a characteristic, namely, $\partial f / \partial \sigma_{\alpha\alpha} = -\partial f / \partial \sigma_{\beta\beta} = 0$, is

$$\frac{\cos 2\phi}{(1-c)} [(\sigma_{\alpha\alpha} - \sigma_{\beta\beta}) \cos 2\phi - 2\sigma_{\alpha\beta} \sin 2\phi] + \sin 2\phi [(\sigma_{\alpha\alpha} - \sigma_{\beta\beta}) \sin 2\phi + 2\sigma_{\alpha\beta} \cos 2\phi] = 0.$$

Solving these two equations for $(\sigma_{\alpha\alpha} - \sigma_{\beta\beta})$ and $\sigma_{\alpha\beta}$,

$$\left. \begin{aligned} \frac{\sigma_{\alpha\beta}}{T} &= (1 - c \sin^2 2\phi)^{\frac{1}{2}}, \\ \frac{(\sigma_{\alpha\alpha} - \sigma_{\beta\beta})}{T} &= \frac{2c \sin 2\phi \cos 2\phi}{(1 - c \sin^2 2\phi)^{\frac{1}{2}}}, \end{aligned} \right\} \quad (15)$$

where the positive square root is to be taken. It may be verified that

$$\frac{d\sigma_{\alpha\beta}}{d\phi} + (\sigma_{\alpha\alpha} - \sigma_{\beta\beta}) = 0.$$

$$\text{Hence} \quad \left. \begin{aligned} \sigma_{\alpha\alpha} &= -p - \frac{1}{2}T \frac{dh}{d\phi}, \quad \sigma_{\beta\beta} = -p + \frac{1}{2}T \frac{dh}{d\phi}, \quad \sigma_{\alpha\beta} = Th, \end{aligned} \right\} \quad (16)$$

where

$$h(\phi) = (1 - c \sin^2 2\phi)^{\frac{1}{2}},$$

and $p = -\frac{1}{2}(\sigma_{\alpha\alpha} + \sigma_{\beta\beta})$ is the mean compressive stress.

To express the relations (12) along the characteristics in terms of the variation of p and ϕ , substitute

$$\sigma_{\xi\xi} = -p + \frac{1}{2}(\sigma_{\alpha\alpha} - \sigma_{\beta\beta}) \cos 2(\phi - \phi_0) - \sigma_{\alpha\beta} \sin 2(\phi - \phi_0),$$

$$\sigma_{\eta\eta} = -p - \frac{1}{2}(\sigma_{\alpha\alpha} - \sigma_{\beta\beta}) \cos 2(\phi - \phi_0) + \sigma_{\alpha\beta} \sin 2(\phi - \phi_0).$$

With the use of (16), equations (12) become

$$\left[\frac{\partial}{\partial \xi} \left\{ p + \frac{1}{2} T \frac{dh}{d\phi} \cos 2(\phi - \phi_0) + Th \sin 2(\phi - \phi_0) \right\} \right]_{\phi=\phi_0} = 0,$$

$$\left[\frac{\partial}{\partial \eta} \left\{ p - \frac{1}{2} T \frac{dh}{d\phi} \cos 2(\phi - \phi_0) - Th \sin 2(\phi - \phi_0) \right\} \right]_{\phi=\phi_0} = 0.$$

Hence

$$\left(\frac{\partial p}{\partial \xi} \right)_P + \left(\frac{1}{2} T \frac{d^2 h}{d\phi^2} + 2Th \right) \left(\frac{d\phi}{\partial \xi} \right)_P = 0,$$

$$\left(\frac{\partial p}{\partial \eta} \right)_P - \left(\frac{1}{2} T \frac{d^2 h}{d\phi^2} + 2Th \right) \left(\frac{d\phi}{\partial \eta} \right)_P = 0.$$

Since P is a general point,

$$\left. \begin{aligned} \frac{p}{2T} + g &= \text{constant on an } \alpha\text{-characteristic,} \\ \frac{p}{2T} - g &= \text{constant on a } \beta\text{-characteristic,} \end{aligned} \right\} \quad (17)$$

where

$$g(\phi) = \int_0^\phi \left(\frac{1}{4} \frac{d^2 h}{d\phi^2} + h \right) d\phi = \frac{1}{4} \left[\frac{dh}{d\phi} \right]_0^\phi + \int_0^\phi h d\phi.$$

By the definition (16) of $h(\phi)$, $g(\phi)$ can be expressed as

$$g(\phi) = -\frac{\frac{1}{2}c \sin 2\phi \cos 2\phi}{(1 - c \sin^2 2\phi)^{\frac{1}{2}}} + \frac{1}{2} E(u, k), \quad (18)$$

where $\sin 2\phi = \text{sn}(u, k)$ and $k^2 = c$.

$\text{sn}(u, k)$ is the Jacobian elliptic function with modulus k (which takes complex values when c is negative), and $E(u, k)$ is the standard elliptic integral of the second kind:*

$$E(u, k) = \int_0^u \text{dn}^2(u, k) du = \int_0^{2\phi} (1 - k^2 \sin^2 \theta)^{\frac{1}{2}} d\theta.$$

For values of ϕ in the range $(-\frac{1}{4}\pi, \frac{1}{4}\pi)$ u lies in the range $(-K, K)$, where K is the quarter-period of the elliptic function. When ϕ lies in the range $(\frac{1}{4}\pi, \frac{3}{4}\pi)$ u must be taken in the range $(K, 2K)$, and so on for other values of ϕ . When $c = 0$, equations (17) reduce to the well-known relations for isotropic metals, due to Hencky (1923):

$$\frac{p}{2T} + \phi = \text{constant on an } \alpha\text{-characteristic,}$$

$$\frac{p}{2T} - \phi = \text{constant on a } \beta\text{-characteristic.}$$

* Tabulated in Jahnke & Emde (1933).

Alternatively, u may be regarded as the independent variable instead of ϕ , and the formulae written in the form

$$\left. \begin{aligned} \frac{\sigma_{\alpha\alpha}}{\sigma_{\beta\beta}} \Big\} &= -p \mp cT \frac{\operatorname{sn} u \operatorname{cn} u}{\operatorname{dn} u}, \quad \sigma_{\alpha\beta} = T \operatorname{dn} u, \\ \frac{\sigma_{\alpha\alpha}}{T} - E(u) &= \text{constant on an } \alpha\text{-characteristic,} \\ \frac{\sigma_{\beta\beta}}{T} + E(u) &= \text{constant on a } \beta\text{-characteristic.} \end{aligned} \right\} \quad (19)$$

To transform (13) in a similar way, substitute

$$v_\xi = v_\alpha \cos(\phi - \phi_0) - v_\beta \sin(\phi - \phi_0),$$

$$v_\eta = v_\alpha \sin(\phi - \phi_0) + v_\beta \cos(\phi - \phi_0),$$

where v_α and v_β are the components of velocity referred to the (α, β) curvilinear axes. Then

$$\left[\frac{\partial}{\partial \xi} \{v_\alpha \cos(\phi - \phi_0) - v_\beta \sin(\phi - \phi_0)\} \right]_{\phi=\phi_0} = 0,$$

$$\left[\frac{\partial}{\partial \eta} \{v_\alpha \sin(\phi - \phi_0) + v_\beta \cos(\phi - \phi_0)\} \right]_{\phi=\phi_0} = 0.$$

$$\text{These reduce to } \left. \begin{aligned} dv_\alpha - v_\beta d\phi &= 0 \text{ along an } \alpha\text{-characteristic,} \\ dv_\beta + v_\alpha d\phi &= 0 \text{ along a } \beta\text{-characteristic.} \end{aligned} \right\} \quad (20)$$

These are identical in form with the equations of Geiringer (1930) for isotropic metals.

5. PROPERTIES OF THE FIELD OF CHARACTERISTICS

In the isotropic theory of plane plastic strain several geometrical properties of the field of characteristics were established by Hencky. Since they are indispensable for the solution of special problems, it is natural to examine whether analogous properties exist in the anisotropic theory.

It is convenient to introduce the quantities (α, β) defined by the relations

$$2\alpha = \frac{p}{2T} - g, \quad 2\beta = \frac{p}{2T} + g. \quad (21)$$

According to (17), α is constant along a β -curve, and β is constant along an α -curve. α and β may therefore be regarded as curvilinear co-ordinates for the field of characteristics. From (21)

$$g = \beta - \alpha, \quad \frac{p}{2T} = \alpha + \beta. \quad (22)$$

Consider any pair of β -curves, $\alpha = \alpha_1$ and $\alpha = \alpha_2$, say. The difference in the values of g where an α -curve ($\beta = \text{constant}$) cuts the two β -curves is, by (22), equal to $\alpha_2 - \alpha_1$, and is therefore independent of β . Thus, the difference in the values of g (or p), where two given curves of one family are cut by any curve of the other family, is a constant. Conversely, any two orthogonal families of curves possessing this pro-

perty constitute a characteristic field for a plastic mass in equilibrium under certain boundary conditions. In Hencky's theorem it is the difference in the values of ϕ that is constant along two given characteristics. This is not true for anisotropic material except in the special case where the difference is $\frac{1}{2}\pi$.

If, now, a section of an α -curve, say, is straight, ϕ is constant along it; hence g , p and α (as well as β) are all constant along the section. It follows from the previous theorem that the corresponding sections of all α -curves are also straight, and so, as a simple consequence, that they are of equal length. An example of such a field is that consisting of radii and concentric circular arcs, whose common centre is a point singularity for the stress distribution. This field appears in the problem solved below.

6. INDENTATION BY A FLAT RIGID DIE

It is evident that the present theory is only applicable throughout a process of plastic deformation so long as the state of anisotropy does not change appreciably, or changes in such a way as to remain uniformly distributed. The indentation of the plane surface of a block of metal by a flat rigid die satisfies the first condition, since the amount of plastic strain in the material near the die is restricted by the surrounding elastic material. The indentation of an isotropic mass was first discussed by Prandtl (1920), and recently by the writer (Hill 1949). In the latter paper it is shown that, if elastic strains are neglected, indentation cannot begin until the plastic region has spread sufficiently far to include the two velocity characteristics extending from some point on the die to the free surface. The development of the plastic zone, as the load on the die is increased, is governed by the state of stress in the elastic, or non-plastic, material. However, with a natural assumption about the general direction in which the plastic zone spreads, the distribution of stress on the die can be determined at the moment when indentation begins, without a detailed knowledge of the stress in the non-plastic material. These considerations are equally applicable to the indenting of an anisotropic mass.

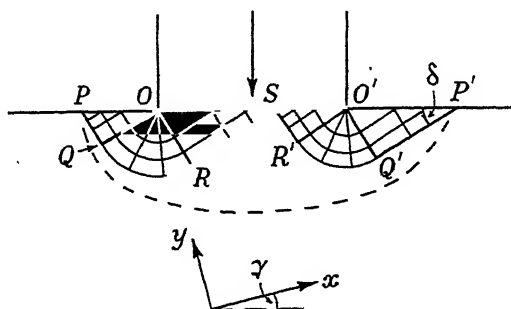


FIGURE 2. Plastic region and characteristics for indenting of block with x axis of anisotropy inclined at an angle γ .

In figure 2 let OO' be the surface of contact (assumed frictionless) between the die and the material. The dimensions of the block, and the length of the die in the direction normal to the plane of the paper, are supposed to be large compared with

the width OO' of the die. At the moment when indentation becomes possible, it is assumed that the plastic zone covers the area between the broken curve and the section PP' of the surface. By the properties of characteristics the state of stress is uniquely determined within the triangles OPQ and $O'P'Q'$ formed by the intersecting pairs of characteristics through O, P and O', P' , respectively. The state of stress in each of these regions is a uniform compression parallel to the surface. Let γ ($0 \leq \gamma \leq \frac{1}{2}\pi$) be the angle between the surface and the x axis of anisotropy (in the sense indicated in figure 2), and let angles POQ and $O'P'Q'$ be denoted by δ . Then, according to (7) and (14),

$$p_0 = T \left(\frac{1-c}{1-c \sin^2 2\gamma} \right)^{\frac{1}{2}}, \quad \delta = \gamma + \frac{1}{2} \cot^{-1} \{ (1-c) \tan 2\gamma \}, \quad (24)$$

where the inverse cotangent is an angle in the interval $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$. It is easy to show that, no matter what the value of γ , δ lies between $\cot^{-1} \sqrt{1-c}$ and $\frac{1}{2}\pi - \cot^{-1} \sqrt{1-c}$.

The characteristic fields in the regions OQR and $O'Q'R'$, defined respectively by the singularities O and O' and the positions of the characteristics OQ and $O'Q'$, consist of radii and concentric circular arcs. The positions of OR and $O'R'$ are determined by the condition that the surface OO' is free from friction. The regions ORS and $O'R'S$ are therefore uniformly stressed, the principal axes of stress being parallel and perpendicular to the surface. Hence the orientation of the characteristics is the same as in regions OPQ and $O'P'Q'$, the angles QOR and $Q'O'R'$ being $\frac{1}{2}\pi$. According to (17) and (18), the value of p in regions ORS and $O'R'S$ is equal to $p_0 + 2TE$, where

$$E = E(K, k) = \int_0^{\frac{1}{2}\pi} (1 - k^2 \sin^2 \theta)^{\frac{1}{2}} d\theta$$

is the complete elliptic integral of the second kind. The pressure on the die is therefore uniformly distributed, and is of amount P , where

$$\frac{T}{2T} = \left(\frac{1-c}{1-c \sin^2 2\gamma} \right)^{\frac{1}{2}} + E. \quad (25)$$

When $c = 0$ this reduces to the Prandtl formula $P = 2T(1 + \frac{1}{2}\pi)$. For a small degree of anisotropy, E can conveniently be calculated from the series

$$E = \frac{\pi}{2} F\left(-\frac{1}{2}, \frac{1}{2}, 1, c\right) = \frac{\pi}{2} \left(1 - \frac{c}{4} - \frac{3c^2}{64} - \frac{5c^3}{256} - \dots \right), \quad (26)$$

where F is the hypergeometric function. The series expansion for P is

$$\frac{P}{2T} = 1 + \frac{\pi}{2} - \frac{c}{2} \left(1 + \frac{\pi}{4} + \sin^2 2\gamma \right) \dots \quad (27)$$

Tables for E as a function of c or k^2 are available (Jahnke & Emde 1933) for the calculation of P for finite degrees of anisotropy (figure 3). If c is positive, P is always less than $2T(1 + \frac{1}{2}\pi)$; if c is negative, P is greater than $2T(1 + \frac{1}{2}\pi)$. Furthermore, P and the configuration of characteristics are the same for orientations γ and $\frac{1}{2}\pi - \gamma$ of the axes of anisotropy; this is due to the symmetry of the anisotropy about directions making 45° with the axes of anisotropy (§ 2).

The position of the point S has so far not been specified. For isotropic material symmetry requires that S should be the midpoint of the die face. For anisotropic material, however, *a priori* considerations cannot decide the position of S , since it is dependent upon the state of stress in the non-plastic material, for it is this that controls the development of the plastic zone.

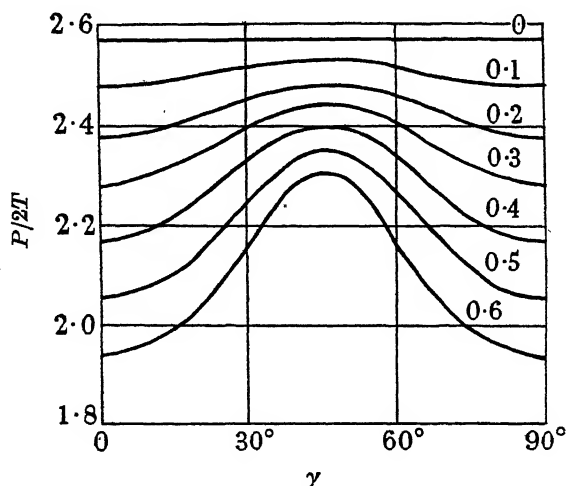


FIGURE 3. Relation between indentation pressure and orientation of axes of anisotropy for various values of c .

As in the isotropic theory, the flow streamlines coincide with the characteristics parallel to $PQRS$ and $P'Q'R'S$, respectively. This follows from the property of equations (20) that the component of velocity in the direction of straight characteristics remains constant along them. The plastic material below $PQRS$ and $P'Q'R'S$, though stressed to the yield limit, is constrained to remain rigid by the surrounding non-plastic material. Hence the constant value of the velocity component along each of the characteristics normal to $PQRS$ and $P'Q'R'S$ must be zero, proving the stated result. It follows that the resultant velocity is of magnitude $V \sec \delta$ in $O'P'Q'R'S$, and of magnitude $V \csc \delta$ in $OPQRS$, where V is the downward speed of the die.

REFERENCES

- Geiringer, H. 1930 *Proc. 3rd Int. Congr. Appl. Mech.* 2, 185.
 Hencky, H. 1923 *Z. angew. Math. Mech.* 3, 241.
 Hill, R. 1948 *Proc. Roy. Soc. A*, 193, 281.
 Hill, R. 1949 *Quart. J. Mech. Appl. Math.* 2, 40.
 Jahnke, E. & Emde, F. 1933 *Funktionentafeln*, 2nd ed. p. 141.
 Prandtl, L. 1920 *Nachr. Ges. Wiss. Göttingen*, p. 74.

A theory of the film phenomena of liquid helium II

By H. N. V. TEMPERLEY, *King's College, University of Cambridge*

(Communicated by D. R. Hartree, F.R.S.—Received 28 February 1949)

A theory of the liquid helium film on the general lines of that due to Schiff (1941) is proposed, the attraction between the walls of the container and the helium atoms being balanced against gravity, but the wave-like nature of the helium atoms and their mutual attractions are now considered. The predictions of the variation of film thickness with height agree with experiment in order of magnitude. A start is made on the problem of calculating the variation of film thickness with temperature, and a new interpretation of the rate of transfer of helium by the film is suggested.

1. INTRODUCTION

Theories of the formation of the Rollin film have been given by Frenkel (1940), Schiff (1941) and by Bijl, de Boer & Michels (1941). The first two authors gave theories which are based on classical considerations, except that they assume an attractive force of London type between the helium atoms and the atoms of the walls. Frenkel (1940) attempts to account for the fact that ordinary liquids do not form such thick films by postulating that the relations between the various surface energies involved may be different in helium from what they are in ordinary liquids, while Schiff (1941) points to the very small viscosity of liquid helium II, which might mean that ordinary liquids are unable to form such films because they would evaporate faster than fresh liquid could flow into them from the bulk liquid, owing to the retarding effect of viscosity. Bijl *et al.* (1941) postulate that the abnormal thickness of the helium film is a consequence of the existence of zero-point energy, which 'blows out' the lattice-spacing of the atoms in the film in much the same way that the lattice-spacing of the atoms in the liquid is greater than can be accounted for on the basis of the interatomic forces (Simon 1934). However, they assume for this zero-point energy of the film the expression $\hbar^2/8mt^2$ per atom, where t is the thickness of the film, which is obtained by neglecting the interactions of the atoms with one another, and assuming that they are described by a wave function of sine type, vanishing at the wall and at the outside boundary of the film. (In what follows, the 'inside' of the film is always to be taken to mean that part nearest the wall.) The form assumed for this zero-point energy has been severely criticized by Mott (1949) on the ground that such a term would be completely absent if interactions were taken into account in any reasonable way. For a film about 100 atoms thick, it certainly seems that any zero-point energy would be mainly settled by the lattice-spacing, which would fix the volume available per atom, rather than directly by the film thickness.

2. DESCRIPTION OF THE THEORY

The energy-levels available to a helium atom moving in an inverse-cube law potential due to the attraction of the wall are first considered. The strength of this attraction can be estimated from the adsorption energy, and it is shown that, by

itself, this field is insufficient to provide bound states for atoms at distances of the order of 3×10^{-6} cm. from the wall. A similar problem is then considered for a helium atom moving in the combined field of the wall, and of layers of atoms covering it, the attraction of the latter being estimated from the interaction potential for helium atoms. The behaviour of such a film in equilibrium with a liquid above and below the λ point is next studied, on the assumption that the liquid undergoes a 'Bose-Einstein condensation', though a detailed knowledge of the energy-levels of the liquid is not required. It is found that the fact that the film is thick only below (and possibly for a short range above) the λ point can be satisfactorily explained. The 'bound' states, representing an atom adsorbed on top of the film, are relatively few in number compared with the 'volume' states, representing an atom moving about in the liquid or vapour phase, but they are, in general, of lower energy. As the film becomes thicker, the energy gain on going into a bound state becomes less, because of the rapidly diminishing attraction of the walls. Below the λ point the film may continue to build up nearly to the thickness at which bound states are no longer possible, but above the λ point the main consideration is that there are relatively few bound states, and they will only fill as long as there is an appreciable energy gain to compensate for the small number of states. A theory of the variation of thickness of film with temperature is given, but there appear to be no reliable data available at present. However, by balancing the change of gravitational energy with height against the change of adsorption energy with thickness, it seems to be possible to get agreement with some unpublished data of L. C. Jackson on the variation of film thickness with height. Lastly, it seems possible to interpret physically the observed 'velocity of transfer' and the observed effect of a constriction in the tube through which the film is flowing.

These bound states are presumably present on any boundary wall, even if it is submerged in the liquid, the bound states being sharply distinguished from the liquid states by the fact that their wave functions vanish except in the immediate neighbourhood of the wall. The existence of these additional states may not have any great effect on the equilibrium properties of the liquid, but certainly will affect the transport properties profoundly. It is hoped to examine this latter point in another paper.

It should be pointed out that another investigation of the writer's (to be published shortly) has shown that existing theories of the 'Bose-Einstein condensation' are unreliable in the transition region, though they give correct results at high and at low temperatures, and the qualitative prediction of a condensation effect is justified by more exact investigation. It is therefore likely that the theory given in this paper of the variation of thickness of film with temperature is unreliable near the transition temperature, but it is put on record pending the discovery of an exact method of handling such theories in the transition region.

3. THE MOTION OF A HELIUM ATOM IN THE ATTRACTIVE FIELD DUE TO THE WALL

It is assumed that the attraction of a wall molecule for a helium atom is of the London inverse sixth-power type, which leads to an inverse-cube law for the

potential of a helium atom near a semi-infinite wall. The corresponding wave equation is

$$\frac{\partial^2 \psi}{\partial x^2} + \left(\frac{8\pi^2 m}{h^2} E + \frac{k^2}{(p+x)^3} \right) \psi = 0, \quad (1)$$

where the closest distance of approach of the atom to the wall is the quantity p , so that ψ must vanish for x zero. If the wall is bare, we take p equal to s , the atomic diameter (not the radius because the radius of an atom of the wall must be allowed for), while if the wall is covered with a film of thickness t , the closest distance of approach is then $s+t$. For E zero, equation (1) can be solved exactly, the solution being $(x+p)^{\frac{1}{2}} J_1 \left(\frac{2k}{(x+p)^{\frac{1}{2}}} \right)$ (the other solution becoming infinite at infinity). We can estimate the constant k^2 , determining the strength of the attractive field, by the following argument. The adsorption energy of helium on glass is known (Keesom & Schweers 1941) to be in the region of 90 cal./mole. On a classical basis this would suggest a value of $\frac{90 \times 4.18 \times 10^7}{6.06 \times 10^{23}} \times (2.6 \times 10^8)^3 = 9.16 \times 10^{-38}$ for $\frac{\hbar^2 k^2}{8\pi^2 m}$ or $k^2 = 1.11 \times 10^{-6}$ if s , the atomic diameter, is taken as 2.6×10^{-8} cm., but we have still to verify that this field is sufficient to give a strongly bound state, so that the actual energy lies near the bottom of the potential well. If p is the closest distance of approach, the wave function must vanish at x zero. The first non-trivial zero of $J_1(q)$ is at $q = 3.83$, so that the condition for a bound state to be possible is $\frac{2k}{p^{\frac{1}{2}}} \geq 3.83$ or, $p \leq 3.02 \times 10^{-7}$ cm. We thus conclude that for $p = s$, the effective diameter of a helium atom (taken as 2.6×10^{-8} cm.), the atom is very strongly bound and our assumption is justified. (As a matter of fact, there is a second bound state lying above the one in question.) It is possible to determine, by a variation method, the height of the bound state above the bottom of the well, and thus to compute a correction to k^2 , but since k^2 is not required very accurately, this will not be done here. For this value of k^2 , the maximum value of p for which a bound state is possible is 3.02×10^{-7} cm. To account for a film of the order of 3×10^{-6} cm. thick, k^2 would have to be increased by a factor of 10, which may be ruled out as quite impossible, in view of the adsorption data.

We therefore conclude that the attraction of the walls is incapable, by itself, of holding films of the observed thickness.

4. ATTRACTION BY THE COMBINED FIELD OF WALL AND HELIUM FILM

The wave equation may now be written

$$\frac{\partial^2 \psi}{\partial x^2} + \left(\frac{8\pi^2 m E}{h^2} + \frac{k^2}{(s+t+x)^3} + \frac{l^2}{(s+x)^3} - \frac{l^2}{(s+t+x)^3} \right) \psi = 0, \quad (2)$$

where the last two terms represent the van der Waals attraction due to the film of thickness t . The value of l^2 is needed as accurately as possible, as the predictions of the theory prove to be sensitive to its precise value. Unfortunately, there are two sources of uncertainty. In the first place, the attractive force between two helium

atoms is not known very accurately, in the second place we have no knowledge of the density of helium atoms in the film. We shall take the value $\frac{1.5 \times 10^{-60}}{r^6}$ erg for the attraction potential between two atoms, which Keesom's discussion (chap. II of his book *Helium*) seems to show to be reasonably consistent with most of the evidence. As we are regarding the atoms as impenetrable spheres, we are not interested in the various repulsion potentials that have been proposed. We shall take the density to be the same as that of the bulk liquid,* the reason for this being that the density of most of the film is probably fixed by the same consideration that fixes that of the bulk liquid, namely, a balance between the attractive forces and the 'zero-point repulsion' (Simon 1934), rather than between the attractive and repulsive forces. No doubt, in the first few inside layers, the density is increased by the effect of the powerful field of the wall to a value approximating to that of the solid, but a rough calculation indicates that this effect can hardly be appreciable beyond the third or fourth layer. (In the first layer the potential energy due to the wall is 90 cal./mole compared with a value of 70 cal./mole for the estimated zero-point energy of the bulk liquid.)

By two simple integrations, we find that the energy of a single atom due to the attraction of a slab of liquid of thickness t is given by the expression

$$\frac{\pi N}{6} \left(\frac{K}{(x+s)^3} - \frac{K}{(x+s+t)^3} \right)$$

if the attraction of a single atom is K/r^6 and the slab contains N atoms/cm.³. Inserting the value 0.146 for the density, we obtain the value 1.73×10^{-38} for $\frac{\hbar^2 l^2}{8\pi^2 m}$ or $l^2 = 2.10 \times 10^{-7}$.

We now determine the binding energy by a variational method. We are interested mainly in the case of small binding energies and $t \gg s$. This suggests that we use the function $e^{-ax}(x+s)^{\frac{1}{2}} J_1\left(\frac{3.83s^{\frac{1}{2}}}{(x+s)^{\frac{1}{2}}}\right)$ as a trial wave function, as it satisfies the boundary conditions at the origin and at infinity, is obviously of the right form e^{-ax} for x large, and is of nearly the right form for x small provided that the binding

energy is small. Inserting this function in the usual formula $E = \frac{\int H\psi\bar{\psi}d\tau}{\int \psi\bar{\psi}d\tau}$ and

minimizing with respect to a , we obtain the following expression for the binding energy when this quantity is small compared with the depth of the potential 'wells':

$$E = -\frac{\hbar^2}{32\pi^2 m} \left[1.19 \left(\frac{16l^2}{(3.83)^4 s^2} - \frac{4}{(3.83)^2 s} \right) + \frac{k^2 - l^2}{t^2} \right]^2 \quad \text{provided } s \ll t. \quad (3)$$

The terms in $1/s^2$ and $1/t^2$ represent respectively the effect of the attractive terms $l^2/(x+s)^3$ and $(k^2 - l^2)/(x+s+t)^3$, while the term in $1/s$ represents the kinetic energy that the atom possesses through being in a bound state. It is interesting to see that

* I am very much indebted to Dr K. R. Atkins for this suggestion.

this formula contains a term in $1/t^2$, strictly speaking it is in $1/(s+t)^2$, just like the de Boer-Michels-Bijl theory, but its origin is entirely different from the $1/t^2$ term in the latter theory.

5. MOTION ALONG THE FILM BOUNDARY: THE EFFECT OF GRAVITY

Since the effect of an inverse-cube law of attraction is not the same as what one deduces on a purely classical basis, we have to examine whether we can use the classical expression for the gravitational energy. We suppose first that gravity acts along the z axis, but that the attractive forces do not vary significantly as we move parallel to the surface. In this case, the wave equation in the z -direction can be solved exactly in terms of Bessel functions of order one-third, the wave function oscillating extremely rapidly with distance as long as the momentum along the z axis is real, but dying away to zero extremely rapidly as soon as the momentum becomes imaginary, so that it is quite safe to use the classical expression for the gravitational energy. If we make the opposite assumption that the attraction of the film does vary periodically with the y and z co-ordinates, so that, at any instant, there are, so to speak, adsorption sites at the surface of the film, then again it is not difficult to show that the classical expression for the gravitational energy is adequate.

The reasonable assumption is made that, travelling up the film, the binding energy of the atoms in the outside layer remains the same, so that the energy mgz has to be balanced against the effect of diminishing t in expression (3). Inserting the values $t = 3.5 \times 10^{-8}$ cm., $l^2 = 2.10 \times 10^{-7}$ cm., $s = 2.6 \times 10^{-8}$ cm., the value for t being taken from Dr L. C. Jackson's unpublished measurements,* we arrive at the following theoretical relation between z and t :

$$-mgz + \frac{\hbar^2}{32\pi^2m} \left[2.05 \times 10^{-7} + \frac{0.90 \times 10^6}{t^2} \right]^2 = \text{constant.} \quad (4)$$

From Jackson's measurements, extending to a range of 1 to 2 cm. above the liquid surface, one can conclude that $dt/dz \sim 10^{-6}$. The spread of his measurements prevents us from determining the second derivative, but it appears to be small and positive. The value of the first derivative deduced from equation (4) is 1.8×10^{-7} , which agrees in order of magnitude with observation, but it is doubtful how much significance can be attached to this, because the theoretical value changes extremely rapidly with the assumed values of s and l^2 because of the subtraction of the two terms in expression (3) involving $1/s^2$ and $1/s$. l^2 is, in turn, sensitive to the assumed spacing of the atoms in the film. Slight disturbing effects due to the attraction of the main body of the liquid are possible very near its surface, but there does not seem to be a very pronounced meniscus in a stationary film. Dr Atkins,* from a consideration of his results on the oscillation of helium films, was led to postulate a law of the simple type $t \propto A/z$, but the constant A is left undetermined on account of the fact that the distribution of velocity through the thickness of a moving film is unknown. A law of the type (4) would conform approximately to a reciprocal law over a limited range of z .

* I am very grateful to Dr Jackson and Dr Atkins for this information.

6. A THEORY OF THE VARIATION OF FILM THICKNESS WITH TEMPERATURE

The constant that appears in equation (4) may be expected to be some function of temperature. At absolute zero, we should expect all bound states whose energies are lower than that of the lowest liquid state to be occupied, and the film to thicken until they are no longer possible. Our imperfect knowledge of the constants involved makes an attempt to improve expression (3) by proceeding to higher approximations, as would be necessary in order to calculate such a limiting thickness, hardly worth while. As the temperature rises, a slight energy defect will be needed to hold the atoms on the surface of the film against thermal agitation, and it is this effect that we want to examine.

We take the following simplified model of the film, supposing that in the y direction it consists of rows of sites of equal energy, each row containing M sites, each of which sites contains r states of equal energy. (This is equivalent to assuming that the variation in the attractive field as we move parallel to the surface of the film is not intense enough to give more than a single, but perhaps r times degenerate, bound state for each site.) We take account of gravity by supposing that the binding energy decreases as z increases, and of the diminishing effect of the attraction of the walls by assuming a decrease in the binding energy as x increases. The energy of a site whose co-ordinates are (x, y, z) is therefore assumed to be

$$E(x, y, z) = \text{const.} + mgz + q(x). \quad (5)$$

We suppose that the states described by equation (5) are in equilibrium with a liquid that is capable of a Bose-Einstein condensation, the fundamental property required being that the lowest energy state in the liquid must be capable of being occupied any number of times up to N , the total number of atoms present. This state may or may not be separated by an energy gap from the excited liquid states, but the distribution of these states must be such as to give the condensation effect at the observed temperature. Now we plainly cannot ascribe to the bound states the property of being occupiable any number of times, because the formation of a monomolecular layer only requires a number of atoms of the order of $N^{\frac{1}{2}}$, so that the formation of a complete film only takes a relatively small number. We therefore make the assumption that each site can only be occupied once, the occupation of one of the r states associated with that site preventing the occupation of any of the others. This assumption is not in any sense equivalent to assuming that the atoms in the film obey Fermi-Dirac statistics, for it is quite compatible with their description by symmetrical wave functions. The states in which one or more sites are occupied twice are not ruled out by any symmetry restrictions, but by the fact that the repulsive fields of the atoms might be expected to give such states very high energies.

It is also obvious that the number of sites available in any one layer must depend on the state of occupation of the layers between it and the wall. We make the simplest possible assumption, namely, that the number of available sites in the y row with co-ordinates (x, z) is just the number of atoms occupying the adjoining y row

$(x - q, z)$, where q is the spacing of the rows, so that a vacant site in any layer implies vacancies in corresponding places in all layers farther away from the wall. We consider the case of r equal to unity, the generalization for other values being easily made. Consider the array of sites obtained by taking a section of the film in an xy plane, so that z is the same for all of these sites. Let M be the number of sites in a y row. Suppose that the number of occupied sites in the rows counting outwards from the wall are N_1, N_2, N_3, \dots , and that the energies associated with the occupation of these sites are E_1, E_2, E_3, \dots . The total number of possible sites, in each of these rows are respectively M, N_1, N_2 , so that the number of possible ways in which an arrangement with these values of the N 's can be realized is $\binom{M}{N_1} \binom{N_1}{N_2} \binom{N_2}{N_3} \dots$, or $\frac{M!}{(M - N_1)! (N_1 - N_2)! \dots}$, which is the coefficient of $a^{N_1} b^{N_2} c^{N_3} \dots$, in the expression $(1 + a + ab + abc + \dots)^M$. A configuration of the type (N_1, N_2, N_3, \dots) contributes $(N_1 + N_2 + N_3 + \dots)$ to the total number of atoms, and $N_1 E_1 + N_2 E_2 + \dots$ to the total energy. The array of sites in the xy plane that we are considering therefore contributes a factor

$$(1 + x\theta^{E_1} + x^2\theta^{E_1+E_2} + x^3\theta^{E_1+E_2+E_3} + \dots)^M \quad (6)$$

to the partition function. Provided that we are in a region of temperature where the saddle-point method of Fowler and Darwin can be applied, the condition that this array should be in equilibrium with the volume states constituting the liquid is simply $\theta = e^{-1/kT}$, $x = \lambda$, where T is the temperature of the liquid and λ is the degeneracy parameter associated with the liquid states. Since the film only contains a relatively small number of atoms, the mode of occupation of the bound states will not have any significant effect on the *equilibrium* properties of the liquid. They will, however, affect the *transport* properties of the liquid profoundly, as they may be expected to occur on any surface to which the liquid has access, even though it is covered by the liquid. It is hoped to study this point in another paper. Here, we are only concerned to obtain a prediction of the way in which film thickness may be expected to vary with temperature, for which purpose we need an estimate of the degeneracy parameter λ for the liquid. In the Bose-Einstein gas model the degeneracy parameter is not significantly different from unity below the transition temperature, but this is only because the energy of the very lowest state is practically zero. The occupation of the lowest state of the liquid varies as $\frac{\lambda e^{-E_0/kT}}{1 - \lambda e^{-E_0/kT}}$, where E_0 is the energy of this state, so that below the transition temperature λ should be practically equal to $e^{E_0/kT}$, while it will fall rapidly from this value as the temperature rises above the transition point.

Examination of the method of deriving expression (6) shows that each of the M factors is the partition function for a single site on the inside layer nearest the wall, together with the sites in the outer layers which do not come into existence until this one is occupied. Thus, the first term corresponds to this site being empty, the second one to just this site being filled, the third term to this site and the one immediately outside it being filled, the fourth term to three sites along a line drawn

outwards in the x direction being filled and so on. The probability of a site in the t th layer, counting outwards, being occupied is thus simply

$$\frac{\sum_{n=t}^{\infty} \lambda^n \exp \left[-\frac{1}{kT} (E_1 + E_2 + \dots + E_n) \right]}{\sum_{n=1}^{\infty} \lambda^n \exp \left[-\frac{1}{kT} (E_1 + E_2 + \dots + E_n) \right]}, \quad (7)$$

the series being continued only until positive values of E_n appear. Unfortunately, the energies E_1, E_2, \dots associated with the addition of extra atoms in each layer cannot be inferred directly from equations (3) and (4), because we have neglected first of all the attractive forces between atoms in the same layer, secondly, the terms arising from $\partial^2 \psi / \partial y^2$, $\partial^2 \psi / \partial z^2$ (which may also be regarded as zero-point energy terms analogous to the term involving $1/s$ in equation (3)), and lastly the attraction of atoms on those nearer the wall, which means that the potential energy of an atom newly added to a layer is not given precisely by equation (4), but that a correction is necessary to allow for the 'loosening' effect of the attraction of this atom on the layers already formed. (Incidentally the last correction is in the right direction to account for the discrepancy between the calculated and observed 'slope' dt/dz of the film.) These points have been taken care of qualitatively by the postulate we made above that the spacing of the atoms in the film is the same as that in the liquid and is due to effectively the same balance between attractive forces and the increase of zero-point energy with density. It therefore seems worth while examining the consequences of various assumptions on these lines.

Assumption A. The energy E_0 lies above the energies E_n associated with the film states. This would result in all the possible bound states being occupied below the λ point so that the thickness of the film should be constant until one goes above this temperature.

Assumption B, that E_0 lies below all the energies E_n associated with the film states, would mean that the film would disappear at absolute zero because all the atoms would go into the state E_0 , so that we seem forced to make

Assumption C, that the energy E_0 lies somewhere within the range of energies E_n . At absolute zero the only liquid state occupied is the state E_0 , and just those film states will be occupied that result in a gain of energy if an atom is transferred from the liquid to the film. As the temperature rises, atoms are excited out of the state E_0 into higher liquid states, and some of the film states of higher energy than E_0 become occupied, that is, we should expect the film to thicken as the temperature rises, becoming thinner as we pass through the transition temperature, that is, the film thickness should pass through a maximum.

As E_0 is probably of the order of -15 cal./mole, that is, numerically very much greater than the changes in energy in most of the film layers due to the effects of the wall and gravity, it is not possible to predict *a priori* which of these three assumptions should be made. In § 7 we shall suggest a little evidence in favour of assumption C.

7. THE 'CHARACTERISTIC VELOCITY' OF THE FILM

Various experiments, such as those by Daunt & Mendelsohn (1939), indicate that the rate of transfer of helium along a surface of given perimeter varies with temperature, but is almost independent of the pressure-head between the two ends of the film. On the picture we have drawn, this process should consist mainly of the jumping of atoms from their existing adsorption sites to neighbouring ones which are empty, and possibly some 'leap-frogging' to empty sites in layers nearer the wall. *Ceteris paribus*, such an effect should be proportional to the minimum perimeter of the surface. The transition probability for such a process could be calculated quantum-mechanically if the height of the potential barriers between neighbouring sites could be estimated, but one would expect it to increase as the film becomes thicker and one recedes from the walls, when the spacing of the atoms probably becomes greater, and the barriers consequently lower. If so, most of the transfer is due to the motion of atoms in the outer layers, and the observed rate of transfer should provide a rough measure of the height of the potential barriers in the outer layers, and thus indirectly a measure of the thickness of the film. Thus, the fact that the rate of transfer seems to pass through a maximum at 1.6 to 1.8° K suggests that the film thickness passes through a maximum here also. The rate of transfer is probably a rather sensitive measure of the thickness, and the fact that it falls to unobservably small values at the transition temperature does not mean that the thickness has fallen to zero, but merely that the potential barriers are preventing the transfer from being rapid. The experiments of Kistemaker (1947) prove that films about 30 atoms thick can be formed even above the transition point.

[*Note added in proof.*] Since the above was written, Dr L. C. Jackson has kindly shown me the results of some preliminary measurements on the variation of film thickness with temperature. These do seem to be consistent with the hypothesis that the thickness passes through a flat maximum somewhere between 1 and 2° K, and thus afford some slight confirmation of the suggestions made in § 7.

I wish to thank the Royal Society for the award of the Smithsonian Research Fellowship, during the tenure of which this work was begun. I also wish to thank Dr L. C. Jackson and Dr K. R. Atkins for communicating their results to me before publication.

REFERENCES

- Bijl, A., de Boer, J. & Michels, A. 1941 *Physica*, 9, 655.
 Daunt, J. G. & Mendelsohn, K. 1939 *Proc. Roy. Soc. A*, 170, 423, 439.
 Frenkel, J. 1940 *J. Phys.* 2, 365.
 Keesom, W. H. & Schweers, J. 1941 *Physica*, 8, 1020.
 Kistemaker, J. 1947 *Physica*, 13, 81.
 Mott, N. F. 1949 *Phil. Mag.* 40, 61.
 Schiff, L. I. 1941 *Phys. Rev.* 59, 838.
 Simon, F. 1934 *Nature*, 133, 529.

Spontaneous emulsification of pure xylene in an aqueous solution through mere adsorption of a detergent in the interface

BY A. KAMINSKI AND J. W. MCBAIN, F.R.S.

Department of Chemistry, Stanford University, Stanford, California

(Received 8 March 1949)

[PLATE 16]

Spontaneous emulsification of a pure liquid can occur. This is demonstrated when liquid hydrocarbons are quietly placed upon the surface of the solution of a suitable detergent. Here special attention has been given to xylene placed upon moderately dilute solutions of dodecylamine hydrochloride.

In many cases violent disruption of the pure liquid occurs, when it is quietly placed upon a soap or detergent solution.

Examination shows that the emulsified droplets still consist of pure solvent stabilized by a coating of adsorbed protective colloid.

The source of the required energy is the energy of adsorption, as well as solubilization of hydrocarbon in the aqueous detergent.

The emulsion formed consists of spherical droplets which therefore retain a positive interfacial tension.

INTRODUCTION

Pure xylene placed upon a detergent solution spontaneously descends into the detergent solution and breaks itself up into fine emulsified droplets whilst retaining a positive interfacial tension. Here a lighter liquid goes of itself into a heavier liquid and enormously increases its surface.

Having observed this phenomenon, we had to find a source for the energy required.

McBain & Woo (1937), in describing instances of true spontaneous emulsification, gave references to the previous literature on this little-known subject, which began with the work of Gad in 1878 and Quincke in 1904.

There have been only two recognized causes of true spontaneous emulsification occurring without any external agitation or force. Many illustrations of each were found by McBain & Woo. The first group is where at the interface a substance is formed which locally reduces the surface tension so markedly and quickly as to produce turbulent motion leading to emulsification. This occurs, for example, when an oil containing a fatty acid is placed on an aqueous solution of alkali, and soap therefore forms at the interface.

A second group was theoretically envisaged by Rashevsky (1928), where a substance is dissolved in one phase but is more soluble in the other. Hence, the dissolved substance diffuses across the interface, but some of the original solvent is carried by collision and viscous drag into the second liquid, where it remains at least temporarily emulsified. Examples of this had already been discovered by Gurwitsch (1913).

The present case is completely different, because no new substance forms at the interface, and nothing diffuses across it.

It may be emphasized that we are using the term 'spontaneous emulsification' in its strict sense, in the absence of even the slightest agitation or external disturbance. This distinguishes it from mere easy emulsification where with low interfacial tension the liquids are easily emulsified by external mechanical means.

While still at Stanford University, Dr J. Vinograd obtained a doubtfully positive result from placing Nujol on an aqueous solution of sodium oleate. Miss L. A. Ehrhardt then discovered, in continuation of such experiments, that pure xylene placed on aqueous solution of dodecylamine hydrochloride spontaneously emulsified. Then in the study of emulsion polymerization at the Shell Development Company, Vinograd, Fong, Ronay & Sawyer (1944), found that some other hydrocarbons exhibit the phenomenon. For example, a lens of ethyl benzene placed upon 5 % aqueous tergitol, or mesitylene placed upon 5 % Intral 231, spontaneously splits up into innumerable emulsion droplets.

QUALITATIVE OBSERVATIONS BY MISS L. A. EHRHARDT

With $N/10$ to $N/2$ dodecyl ammonium chloride a uniform layer of emulsion forms at the interface and rapidly penetrates the aqueous phase to a depth of 1 or more cm. This can occur within a few seconds. On standing for weeks or months, the emulsion in $N/2$ dodecylamine hydrochloride may change into a rigid anisotropic liquid crystal, with a clear aqueous layer below and a clear layer of excess xylene above.

THE PRESENT INVESTIGATION

Hydrocarbons other than xylene

Benzene, toluene, mesitylene and cyclohexane produced spontaneous emulsions on the surface of a $N/5$ dodecylamine hydrochloride solution. The rate of emulsification and the extent of the emulsion layer formed decreased in the following order: cyclohexane, benzene, toluene, xylene, mesitylene (this is also the order of decreasing solubility and also of decreasing solubilization). The stability of the emulsion layers formed was greatest for mesitylene and least for cyclohexane. With benzene and toluene the emulsion layer penetrates much deeper into the aqueous phase than with xylene, but has a much greater tendency to change into a rigid gel.

All the remaining experiments were carried out with xylene (Kahlbaum's purest) and dodecylamine hydrochloride supplied by Armour and Company.

Hydrocarbons in other detergents

Spontaneous emulsions have been observed to form in aqueous solutions of several other detergents, all of which are good solubilizing agents. Cetyltrimethyl ammonium bromide, Detergent 'X' (a non-ionic polymer of ethylene oxide with an iso-octyl phenol), and Triton X-100 (similar and also non-ionic) gave violent emulsification with xylene and cyclohexane. Emulsification proceeded much more

violently in aqueous solutions of Detergent 'X' and Triton X-100 than in dodecylamine hydrochloride. In the solution of cetyltrimethyl ammonium bromide, emulsification of xylene was much slower. M. H. McHan at Stanford University had noted that toluene emulsifies in aqueous solutions of Aresklene.

Immobilization of aqueous layer by gelatin

To avoid possibility of stirring and convection, $N/5$ dodecylamine hydrochloride solution was completely rigidified by the addition of gelatin. When xylene was placed upon it, the emulsification nevertheless proceeded, but at a much slower rate. It required 4 days to produce an emulsified layer comparable in depth to that which forms in the ordinary liquid solution in a quarter of an hour. However, the emulsification process did not terminate at this point. Columns of emulsion started to shoot out of the emulsion layer and went down to the very bottom of the container. After 10 days the whole gelatinized solution had become turbid. The presence of emulsion droplets in the rigid jelly was confirmed by microscope examination.

Effect of concentration of the detergent

Spontaneous emulsification is not noticeable in solutions of dodecylamine hydrochloride less than decinormal.

With decinormal detergent, the emulsification is very slow, requiring hours or days for completion. Moreover, the emulsion layer penetrates downwards only 0.5 to 2 mm. The first signs of emulsification appear about 30 min. after contact with the xylene in the form of a very faint, hardly perceptible boundary at a distance of 0.5 to 2 mm. below the interface. The boundary becomes more evident, and a milky white layer of emulsion begins to form in the space between this boundary and the interface. The emulsification continues more and more slowly for at least 1 day until a viscous layer of emulsion has formed. The amount of emulsion formed is independent of the amount or excess of xylene, provided the whole surface is covered. With a microscope it can be seen that drops of emulsion shoot out of the interface, travel a certain depth down into the aqueous phase and then return up into the xylene. It is during the downward motion of the larger drops that smaller ones detach themselves and are left to form the opaque emulsion layer. Seen from above the emulsion is white and opaque, but at a glancing angle the interface presents a mirror surface.

With $N/5$ detergent the emulsion layer is much deeper and it forms in a few minutes. Emulsification proceeds as long as the aqueous phase is in contact with the xylene and is not separated from it by a viscous emulsion layer.

The process of emulsification of a drop of xylene on $N/3$ dodecylamine hydrochloride is shown in figure 1. It illustrates the turbulent flow set up in the detergent solution. The flow is outwards from the drop of xylene, producing a rising column of the solution up to the drop, where the emulsion is cast off.

With $N/2$ dodecylamine hydrochloride the whole process is much more violent and is completed in a few seconds. Even with 0.4N solution, this rigid layer begins to form after a few days.

Figure 2, plate 16, is a photographic record of the emulsification that takes place when xylene is placed upon four different concentrations of solutions, each in a narrow cylinder. Each experiment is in duplicate, one, *a*, with a few drops of xylene covering only a small fraction of the surface, and the other, *b*, with a column of xylene above the aqueous liquid.

The photographs illustrate vividly the pronounced effect which the concentration of the detergent solution exerts upon the rate and intensity of the emulsification process. With a $N/2$ or a $N/4$ dodecylamine hydrochloride solution, the formation of the emulsion is noticeable at the instant the xylene is deposited on the surface. In solutions which are less than $N/8$ the emulsion layer is perceptible only after several hours, and it is completely formed only after several days.

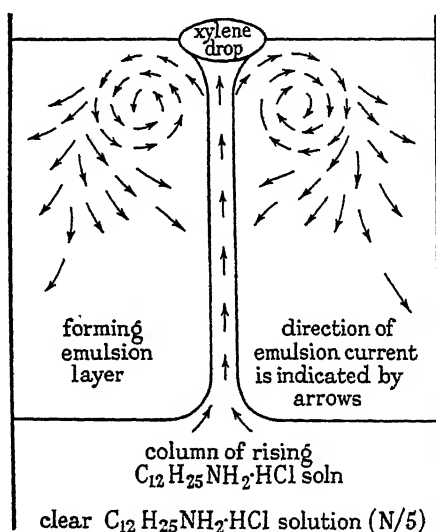


FIGURE 1. Diagrammatic representation in cross-section of the convection currents set up when a drop of pure xylene is placed upon a solution of dodecyl ammonium chloride and begins to emulsify.

Spontaneous emulsification from a small lens of pure xylene on an extended surface

The foregoing described examples in which a layer of xylene was placed upon a layer of aqueous solution in a cylindrical vessel. Other observations were carried out on a drop of xylene placed upon a large free surface of aqueous solution.

Fate of a small drop of hydrocarbon placed upon the surface of an aqueous detergent

When a drop of hydrocarbon or other only slightly soluble organic liquid is placed upon the surface of an aqueous solution of soap, bile salt, or detergent, it need not necessarily emulsify. It will be solubilized if its molecular weight is not too great. In the simplest cases as with styrene, placed upon a soap solution (Vinograd *et al.* 1944), the droplet of styrene will retain the shape of a round lens whilst gradually shrinking until it vanishes, having been completely solubilized. No emulsion is formed unless external agitation is employed and excess of styrene is present.

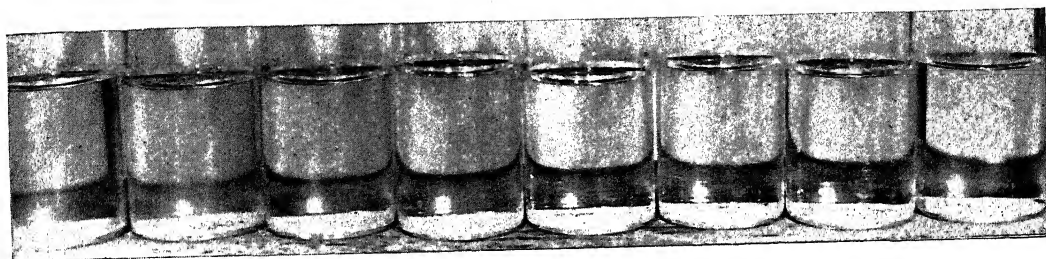
0.45 N
a b

0.23 N
a b

0.11 N
a b

0.06 N
a b

before adding xylene



1

2

3

4

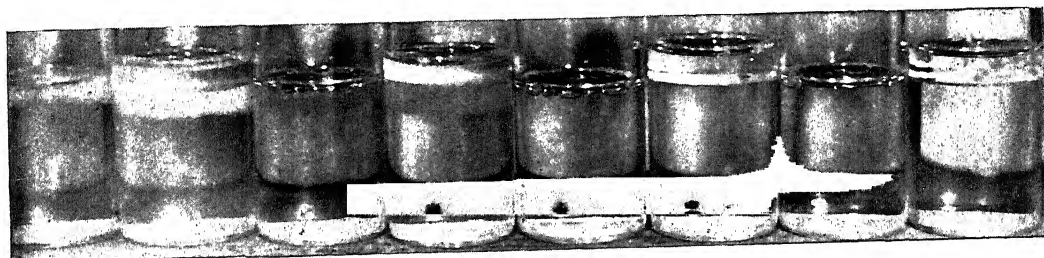
5

6

7

8

after 3 minutes



1

2

3

4

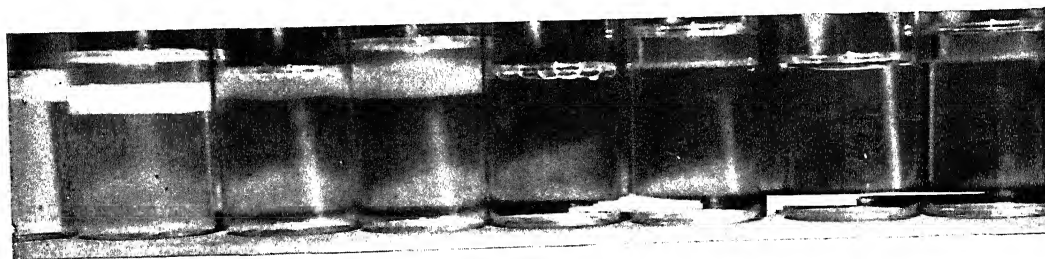
5

6

7

8

after 13 minutes



1

2

3

4

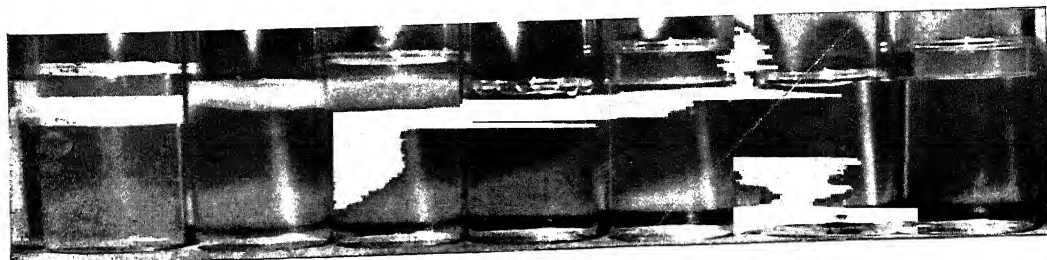
5

6

7

8

after 50 minutes



1

2

3

4

5

6

7

8

FIGURE 2. Photographs of spontaneous emulsification of xylene placed upon four different concentrations of dodecyl ammonium chloride solutions: (a) when a small lens is used, and (b) when a complete layer of xylene is employed.

However, in the interesting cases discussed here, the droplet may seem to explode in a violent manner to form the emulsion. Alternatively, its boundaries may break very irregularly and shoot off a multitude of emulsion droplets.

The concentration of the detergent solution has an effect not only on the amount of emulsion formed, and the rate of emulsification, but also upon the form which the dispersed phase takes. If xylene forms a spontaneous emulsion in a dilute dodecylamine hydrochloride solution (about $N/10$), no disruption of the interface can be observed under a microscope, except in the case of drops of at least several millimetres in diameter. These break up into smaller drops. The small emulsion droplets which are eventually formed are perfectly spherical and average about 50μ in diameter.

In more concentrated solutions of dodecylamine hydrochloride ($N/5$ to $N/2$) the interface between the aqueous phase and the xylene is broken up in many places, and the xylene streams through the holes into the water phase only to form a new droplet a certain distance from the old interface. The new drops formed in this manner are often irregularly shaped; and with very concentrated solutions ($0.8N$) the drops assume myeline forms.

In the cases described, the lens does not spread on the aqueous surface to form a uniform visible film, and, under the circumstances, it is difficult to ascertain whether it attempts to form a monolayer. Burdon (1940) describes the behaviour of a drop of the soluble substance, aniline, on pure water. Here the same violent commotion is displayed. But if droplets of emulsion size are thus injected into the water, they are not apparent because they dissolve. If the water is already saturated, the aniline drop remains at rest.

Temperature and stability of the emulsions

The effect of temperature has not been fully explored. However, it may determine whether or not a particular liquid will emulsify spontaneously. In the present case with xylene on dodecylamine hydrochloride solution there is no spontaneous emulsification above 30°C , and an emulsion made by agitation tends to break down. At very low temperatures the emulsion is again unstable. The emulsion is more stable above 15°C , but the rate of formation is less than at room temperature. At room temperature the emulsions are extremely stable and can be broken up only with difficulty, in a powerful centrifuge.

DISCUSSION

The main factor that produces spontaneous emulsification is the turbulence caused by temporary inequalities of interfacial tension. Eddy currents then tear off droplets of the oil. Since the detergents under discussion are all good protective or stabilizing agents, any small oil droplets will remain suspended as an emulsion.

There is a parallelism between spontaneous emulsification and solubilization, and probably also solubility. The oil begins to leave the continuous oil phase, partly as a result of its tendency to be solubilized by the aqueous detergent. Moreover, the detergent, due to adsorption, is much more concentrated in the initial

interface, and the solubilized complex is known to be formed with positive affinity, that is, with decrease of free energy (McBain & McBain 1936; O'Connor & McBain 1940).

Source of the energy causing spontaneous emulsification of a pure liquid

One source of the energy required to submerge a light liquid and enlarge its interface was determined by the results of a quantitative analysis for each constituent of the system before and after emulsification.

Data for quantitative determination of each component before and after spontaneous emulsification to form three layers: xylene, emulsion, clear aqueous layer underneath

A weighed amount of 0.23-N dodecylamine hydrochloride solution was introduced into a separating funnel and a known amount of *p*-xylene was deposited on the surface. The emulsification process was allowed to proceed undisturbed for 24 hr. Clear xylene was next removed with a hypodermic syringe whose needle was bent at the end. The clear dodecylamine hydrochloride solution was drawn off from the bottom and the emulsion layer was washed out with acetone. Two such operations were necessary to complete the analysis: one for the determination of the amount of hydrochloric acid in the aqueous phase, and another for the determination of the total solids content of the aqueous phase. The amount of hydrochloric acid present was determined by potentiometric titration. The solid content was obtained from the weight of the residue left on evaporating the aqueous phase to dryness with an excess of hydrochloric acid. (The evaporation was carried out by heating the sample in a Petri dish for several hours under a porcelain radiating plate.) The data are tabulated in table 1.

The result of the quantitative examination is to show that the xylene remains pure xylene; the aqueous solution of detergent becomes more dilute; the dodecylamine and hydrochloride are abstracted into the interface in exactly equivalent amounts leaving the solution of the same pH as before. Hence, unchanged detergent is adsorbed in the interface.

Concurrently, xylene is being solubilized in the colloidal particles of the detergent. For this to occur, xylene molecules must move across the interface. It has been shown (McBain *et al.* 1937) that moving molecules carry with them other neighbouring molecules and can even carry an excess of other molecules into a second phase, which would finally reject them. Here such rejection would take the form of emulsification.

Another somewhat similar case of spontaneous emulsification, observed but not studied, is that indicated by Cofman (1929) only in the title of figure 13 of his article. We confirm that water placed upon a 10 % solution of ferric chloride ($\text{FeCl}_3 \cdot 6\text{H}_2\text{O}$) in nitrobenzene exhibits turbulence at the interface and produces an emulsion of water in the heavier liquid, leaving a clear layer of water on top.

When the detergent solution is already saturated with the hydrocarbon (by solution and solubilization); liquid hydrocarbon placed thereon does not spontaneously emulsify.

Previous investigators have recorded a high positive temperature coefficient for solubilization, but it is to be noted that their measurements referred only to dyes.

TABLE 1. QUANTITATIVE ANALYSIS OF ORIGINAL; AND FINAL DISTRIBUTION OF COMPONENTS AMONG THE PHASES PRESENT

phase	weight of components (g.)					pH
	(1) xylene	(2) water	(3) dodecyl- amine	(4) HCl	(5) total (for each phase)	
original composition						
(A) xylene	12.013	—	—	—	12.013	2.1
(B) aqueous	—	73.422	3.057	0.654	77.133	
total	12.013	73.422	3.057	0.654	89.146	
final compositions						
(C) xylene	10.199	0.001	0	0	10.200	2-2.1
(D) emulsion	1.464	6.299	0.433	0.087	8.283	
(E) aqueous	0.350	67.122	2.621	0.567	70.660	
total	12.013	73.422	3.054	0.654	89.143	

Calculations and sources of results

- A (1) 12.013 g. weighed directly.
 B (2) $77.133 - 3.711 = 73.422$ g. $3.711 = (0.0481)(77.133)$. 0.0481 is the ratio of solid to liquid content in the solution.
 B (3) $3.711 - 0.634 = 3.057$ g.
 B (4) $(0.233)(77.1/1000)(36.5) = 654$ g.
 C (1) 10.199 g. obtained by direct weighing.
 C (2) 0.001 g. obtained from the *International Critical Tables*.
 C (3) 0. Evaporation of the extracted xylene layer did not leave any residue.
 C (4) 0. If any xylene layer is extracted from the system after the emulsification process has reached equilibrium and shaken with some distilled water, the pH of the water does not change.
 C (5) 10.200 g. obtained by addition.
 D (1) $12.013 - 10.199 - 0.350 = 1.464$ g.
 D (2) $73.422 - 67.122 - 0.001 = 6.299$ g.
 D (3) $0.520 - 0.087 = 0.433$ g.
 D (4) $0.654 - 0.567 = 0.087$ g.
 D (5) 8.283 g. obtained by addition.
 E (1) $(0.50)(70/100) = 0.350$ g. 0.50 g. is the amount of xylene solubilized by 100 g. of the dodecylamine hydrochloride solution.
 E (2) $70.660 - 3.188 - 0.350 = 67.122$ g. 3.188 g. is the weight of the solid residue, remaining after the extracted aqueous phase has been evaporated with an excess of HCl.
 E (3) $3.188 - 0.567 = 2.621$ g.
 E (4) $(0.220)(70.66/1000)(36.46) = 0.567$ g. 0.220 g. is the amount of equivalents of HCl per 1000 g. of solution remaining in the aqueous phase after the emulsification process has reached equilibrium.
 E (5) 70.660 g. weighed directly.

Conclusions

- (a) The xylene phase remains unchanged by the emulsification process.
 (b) The dodecylamine hydrochloride is adsorbed as such from the aqueous solution into the interfaces of the emulsion of pure xylene, leaving the pH of the aqueous layer unchanged.

In the Stanford Research Institute our collaborators, Dr A. P. Brady and Mrs H. Huff (née McHan), find that the temperature coefficient of solubilization of a liquid hydrocarbon is strongly negative. We find this to be especially the case in our experiments with xylene in dodecylamine hydrochloride; the slightest warming of the saturated solution produces turbidity.

A new phenomenon was also discovered by Brady and Mrs Huff. If benzene is added to a clear aqueous solution of dodecylamine hydrochloride, the solution remains clear until it is apparently saturated with benzene; then it becomes turbid. Surprisingly, further addition of benzene clears the solution again until it is finally saturated. These two zones of clarity exist only in the neighbourhood of room temperature and not above about 35° C. The nature of the intermediate turbidity remains to be determined. The phenomena are reversible.

REFERENCES

- Burdon, R. S. 1940 *Surface tension and the spreading of liquids*, p. 64. Cambridge Physical Tracts. Cambridge University Press.
 Cofman, V. 1929 *J. Chem. Educ.* 6, 1095.
 Gad, J. 1878 *Arch. Anat. Physiol., Lpz.*, p. 181.
 Gurwitsch, L. 1913 *Wissenschaftliche Grundlagen der Erdölbearbeitung*, p. 200; translation by Moore (1932), p. 430. London: Chapman and Hall, Ltd.
 McBain, J. W. & McBain, E. L. 1936 *J. Amer. Chem. Soc.* 58, 2610.
 McBain, J. W. & Woo, Ts-Ming 1937 *Proc. Roy. Soc. A*, 163, 182.
 O'Connor, J. J. & McBain, J. W. 1940 *J. Amer. Chem. Soc.* 62, 2855.
 Quincke, G. 1904 *Ann. Phys., Lpz.*, 15, 1.
 Rashevsky, N. 1928 *Z. Phys.* 46, 568.
 Vinograd, J. R., Fong, L. L., Ronay, G. S. & Sawyer, W. M. 1944 *Studies in emulsion polymerization*. Presented before the Division of Colloid Chemistry, 108th meeting of the American Chemical Society, New York, 13 September.

The diffraction of blast. I

By M. J. LIGHTHILL, *Department of Mathematics, University of Manchester*

(Communicated by S. Goldstein, F.R.S.—Received 25 November 1948)

The behaviour of a plane shock, of any strength; travelling along a wall, when it reaches a corner where the wall turns through a small angle δ , is investigated mathematically by use of a linearized theory of anisentropic flow (§§ 2 to 5). At a convex corner pure diffraction occurs; at a concave corner Mach reflexion (§ 8). The shape of the shock (§ 6) and the pressure distribution over the wall (§ 7) are calculated for a variety of shock Mach numbers from 1 to ∞ . The connexion, for weak shocks, with acoustic theory, is displayed (§ 9). The work will be used in later parts to assist in a hypothetical description of the flow when δ is not small.

1. INTRODUCTION

Let a plane shock of any strength move uniformly into still air, which is bounded by a plane wall perpendicular to the shock. Suppose that at a certain line (parallel to the shock) this wall joins another plane wall so that the two form a corner convex to the flow, of angle say $\pi - \delta$. When the shock reaches this line it will be diffracted and thereafter will not be plane, being weaker near the wall than elsewhere. In using mathematical physics to predict what will happen, viscous stresses and conduction of heat can reasonably be neglected, since neither will have time to produce sub-

stantial effects during the extremely rapid passage of the shock. In consequence the only physical constants defining the problem will be U , the original shock velocity p_0 and ρ_0 , the pressure and density in still air, and δ . Since these cannot be combined to produce a fundamental length- or time-scale, it follows that while, in this two-dimensional problem, each physical quantity is a function of the two space co-ordinates X , Y (with the corner as origin) and the time t , these variables can only occur therein in the combinations X/t and Y/t . In this part it will be assumed that δ , and hence also the variations in velocity and pressure behind the shock, are small. By taking δ negative the behaviour of the shock at a concave corner can also be studied. Previous work has been confined to the case of a weak shock (Sommerfeld 1901).

2. MATHEMATICAL FORMULATION

Let the velocity, pressure, and density behind the shock before it reaches the corner, be q_1 , p_1 and ρ_1 ; then by conservation of mass, momentum and energy at the shock,

$$\rho_1(U - q_1) = \rho_0 U, \quad \rho_0 U q_1 = p_1 - p_0, \quad \rho_0 U \left(\frac{1}{2} q_1^2 + \frac{5p_1}{2\rho_1} - \frac{5p_0}{2\rho_0} \right) = p_1 q_1, \quad (1)$$

assuming that air is a perfect gas with adiabatic index $\gamma = 1.4$ (so that $1/(\gamma - 1) = \frac{5}{2}$), which is a reasonable approximation in this problem for shock pressure ratios up to about 30 (for values in excess of which the problem is of much smaller practical importance).

Equations (1) solved for q_1 , p_1 , ρ_1 give

$$q_1 = \frac{5}{6} U \left(1 - \frac{a_0^2}{U^2} \right), \quad p_1 = \frac{5}{6} \rho_0 (U^2 - \frac{1}{7} a_0^2), \quad \rho_1 = 6\rho_0 / \left(1 + \frac{5a_0^2}{U^2} \right), \quad (2)$$

where $a_0 = (7p_0/5\rho_0)^{\frac{1}{2}}$ is the velocity of sound in still air. Let M be the Mach number of the shock, U/a_0 , which must exceed unity, and let $M_1 = q_1/a_1$ be the Mach number of the uniform flow behind the shock, so that

$$M_1 = q_1(7p_1/5\rho_1)^{-\frac{1}{2}} = \frac{5(M^2 - 1)}{[(7M^2 - 1)(M^2 + 5)]^{\frac{1}{2}}}. \quad (3)$$

Then this flow is supersonic ($M_1 > 1$) when $M > (7 + \sqrt{34})^{\frac{1}{2}} 3^{-\frac{1}{2}} = 2.068 \dots$, and subsonic when $M < 2.068 \dots$; these cases will frequently need to be distinguished as the 'supersonic case' and the 'subsonic case'.

After diffraction let the velocity, pressure, density and entropy at any point be q_2 , p_2 , ρ_2 , s_2 . Choose (X, Y) axes with origin at the corner and X -axis along the original wall produced. If $D/Dt = \partial/\partial t + \mathbf{q}_2 \cdot \nabla$ signifies time-rate-of-change for a given fluid element, the equations of conservation of mass and momentum can be written

$$\frac{D\rho_2}{Dt} + \rho_2 \operatorname{div} \mathbf{q}_2 = 0, \quad \frac{D\mathbf{q}_2}{Dt} + \frac{1}{\rho_2} \nabla p_2 = 0; \quad (4)$$

and if there is no heat transfer between fluid elements by friction conduction or radiation the entropy will satisfy $Ds_2/Dt = 0$.

On the assumption that q_2 , p_2 , ρ_2 differ only by small quantities from the values $(q_1, 0)$, p_1 , ρ_1 which they had before diffraction, the equations of motion can be approximated as

$$\frac{\partial \rho_2}{\partial t} + q_1 \frac{\partial \rho_2}{\partial X} + \rho_1 \operatorname{div} \mathbf{q}_2 = 0, \quad \frac{\partial \mathbf{q}_2}{\partial t} + q_1 \frac{\partial \mathbf{q}_2}{\partial X} + \frac{1}{\rho_1} \nabla p_2 = 0, \quad \frac{\partial s_2}{\partial t} + q_1 \frac{\partial s_2}{\partial X} = 0. \quad (5)$$

The entropy and density variations can now be eliminated from the entire problem, since by virtue of the last equation $\partial \rho_2 / \partial t + q_1 \partial \rho_2 / \partial X$ can be replaced by

$$\left(\frac{\partial \rho_1}{\partial p_1} \right)_s \left(\frac{\partial p_2}{\partial t} + q_1 \frac{\partial p_2}{\partial X} \right) = \frac{1}{a_1^2} \left(\frac{\partial p_2}{\partial t} + q_1 \frac{\partial p_2}{\partial X} \right), \quad (6)$$

owing to the thermodynamic principle that density is a function of pressure and of entropy alone.

If now the transformations

$$\frac{X - q_1 t}{a_1 t} = x, \quad \frac{Y}{a_1 t} = y, \quad \frac{\mathbf{q}_2}{q_1} = (1 + u, v), \quad \frac{p_2 - p_1}{a_1 q_1 \rho_1} = p \quad (7)$$

be made, and the fact, shown in § 1, that u, v, p depend only on x and y , be used, then the first two equations of (5) become

$$x \frac{\partial p}{\partial x} + y \frac{\partial p}{\partial y} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}, \quad x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = \frac{\partial p}{\partial x}, \quad x \frac{\partial v}{\partial x} + y \frac{\partial v}{\partial y} = \frac{\partial p}{\partial y}. \quad (8)$$

A word is necessary on the position of the axes for which the equations of motion take the simple form (8). The origin is at a point on the original wall produced. The part of the shock which is still straight lies along a fixed line

$$x = \frac{U - q_1}{a_1} = M_1 \left(\frac{U}{q_1} - 1 \right) = \left(\frac{M^2 + 5}{7M^2 - 1} \right)^{\frac{1}{2}} = k < 1. \quad (9)$$

The corner is at $(-M_1, 0)$.

The conditions at a point immediately behind the diffracted shock will depend on the local velocity of the shock normal to itself; they will be given by the right-hand sides of equation (2) if U be replaced therein by this velocity, the direction of flow behind the shock being normal to it. Now in the (X, Y) system of co-ordinates each point on the shock moves with velocity vector $(X/t, Y/t)$, since the whole field suffers a uniform expansion in time about the corner. Hence the velocity of the shock normal to itself at any point is \mathbf{h} , where $t\mathbf{h}$ is the vector perpendicular drawn from the corner to the tangent to the shock at that point. In terms of \mathbf{h} the boundary conditions at the shock are

$$\mathbf{q}_2 = \frac{5}{6} \mathbf{h} \left(1 - \frac{a_0^2}{h^2} \right), \quad p_2 = \frac{5}{6} \rho_0 \left(h^2 - \frac{1}{7} a_0^2 \right). \quad (10)$$

Let the equation of the shock in the new co-ordinates be $x = k + f(y)$, then $f(y)$ will presumably be uniformly small if δ is small. On this assumption of a nearly straight shock

$$t\mathbf{h} = (X - Y dX/dY, -X dX/dY); \quad (11)$$

but $X = a_1 tx + q_1 t = Ut + a_1 tf(y)$ and $Y = a_1 ty$; hence \mathbf{h} can be taken as

$$\mathbf{h} = (U + a_1 f(y) - a_1 y f'(y), -U f'(y)). \quad (12)$$

Hence, combining (7), (10) and (12), the approximate shock boundary condition is deduced that, on $x = k$,

$$u = a_1[f(y) - yf'(y)] \frac{d}{dU} \log \left[\frac{5}{6} U \left(1 - \frac{a_0^2}{U^2} \right) \right] = \frac{a_1}{U} [f(y) - yf'(y)] \frac{M^2 + 1}{M^2 - 1},$$

$$v = -f'(y), \quad p = \frac{p_1}{a_1 q_1 \rho_1} a_1 [f(y) - yf'(y)] \frac{d}{dU} \log (U^2 - \frac{1}{7} a_0^2)$$

$$= \frac{p_1}{q_1 \rho_1} [f(y) - yf'(y)] \frac{2U}{U^2 - \frac{1}{7} a_0^2}. \quad (13)$$

Hence, on $x = k$, $u = Ap$, $y dv / \partial y = B \partial p / \partial y$, (14)

where $A = \frac{a_1 q_1 \rho_1 (M^2 - \frac{1}{7}) (M^2 + 1)}{2 p_1 M^2 (M^2 - 1)} = \frac{M^2 + 1}{2 M^2} \left(\frac{7 M^2 - 1}{M^2 + 5} \right)^{\frac{1}{2}}$,

$$B = \frac{\rho_1 q_1}{p_1} \frac{U^2 - \frac{1}{7} a_0^2}{2U} = \frac{3(M^2 - 1)}{M^2 + 5}. \quad (15)$$

Conversely, if equations (14) hold, there is a function $f(y)$ satisfying (13); hence equations (14) may be taken as the complete shock boundary condition.

The problem is now reduced to mathematical terms. The equations (8) must be solved under the boundary conditions that, on $y = 0$, $v = -\delta$ for $x > -M_1$ and $v = 0$ for $x < -M_1$; that on $x = k$ equations (14) hold, and that on the remaining boundary between the uniform flow and the disturbed flow $u = v = p = 0$.

3. ELIMINATION OF u AND v

Now equations (8) occur also in a problem which recently has received much attention, following Busemann (1943). This is that of steady supersonic flow, in which departures from uniformity are small, entropy variations are neglected, and the velocities are constant on all straight lines through the origin (the so-called cone-field problem). If here u , v are the disturbance velocities in the x , y directions (perpendicular to the main stream), where x , y and also p , the pressure, are suitably altered in scale, then equations (8) hold, together (however) with the condition of irrotationality $\partial v / \partial x = \partial u / \partial y$ which is absent in the problem of § 2.

An immediate consequence of equations (8) is that

$$\nabla^2 p = \left(x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} + 1 \right) \left(x \frac{\partial p}{\partial x} + y \frac{\partial p}{\partial y} \right), \quad (16)$$

a second-order equation in p . When the condition of irrotationality holds, the same equation is satisfied by u and by v . But in its absence u and v satisfy third-order equations. This indicates that the present problem should be solved in terms of p alone (which in any case is the most interesting variable), by eliminating u and v .

As Busemann observed, equation (16) is hyperbolic for $x^2 + y^2 > 1$ and elliptic for $x^2 + y^2 < 1$; its characteristics are all the tangents to the circle $x^2 + y^2 = 1$; no other line, except an arc of the circle itself, can be part of a boundary between a region where some solution of (16) is constant and one where it is not. It is reasonable to

assume that the region of non-uniform flow (behind the shock) is the smallest region consistent with this fact and with the fact that the corner is either inside it or on its boundary. This means that in the subsonic case (when the corner $(-M_1, 0)$ is inside the unit circle) this boundary is an arc of the circle, and that in the supersonic case it is the tangent from the corner to the circle plus an arc of the circle. (We shall find that this hypothesis yields a solution; had it not, it would have been necessary to allow a larger disturbed region.)

On the x -axis v is a step-function, and $\partial v/\partial x = 0$ except at the corner where it does not exist; also $y = 0$; hence by (8) $\partial p/\partial y = 0$.

On the part of the circle $x^2 + y^2 = 1$ with $y > 0$ and $x < k$, in the subsonic case (when this constitutes the whole 'third boundary') $p = 0$; but in the supersonic case this is only true beyond the point of contact of the tangent from $(-M_1, 0)$, i.e. for $x > -M_1^{-1}$. In the region between the tangent and the circle disturbances are propagated along characteristics; but the only disturbance is a concentrated one emanating from the corner; hence the flow in this region is a uniform one parallel to the wall, separated from the previous uniform flow by a discontinuous expansion, which is of course the approximation given by the linearized theory to a Prandtl-Meyer expansion. (Steady flow obtains in this region; the only 'unsteadiness' associated with the region is the fact that it is constantly growing.) Hence on the unit circle, for $x < -M_1^{-1}$, the pressure has the value given by a linear approximation to the Prandtl-Meyer expansion at the corner, i.e. $p_2 = p_1 - \rho_1 q_1^2 \delta (M_1^2 - 1)^{-\frac{1}{2}}$, or

$$p = -M_1 \delta (M_1^2 - 1)^{-\frac{1}{2}}. \quad (17)$$

Thus the boundary values of p on the circle are discontinuous in the supersonic case. The corresponding property in the subsonic case is that, at the corner ($x = -M_1$), though $(\partial p/\partial y)_{y=0} = 0$ on both sides,

$$\lim_{y \rightarrow 0} \int_{-M_1-c}^{-M_1+c} \frac{\partial p}{\partial y} dx = -M_1 \int_{-M_1-c}^{-M_1+c} \frac{\partial v}{\partial x} dx = M_1 \delta, \quad (18)$$

using equations (8) and the fact that v increases by $-\delta$ at the corner.

From the shock boundary conditions (14) and equations (8) only one condition on p can be deduced, namely, that

$$k \left(k \frac{\partial p}{\partial x} + y \frac{\partial p}{\partial y} \right) = \frac{\partial p}{\partial x} - A y \frac{\partial p}{\partial y} + B k y^{-1} \frac{\partial p}{\partial y} \quad (19)$$

on $x = k$. (The intermediate step is that both sides are equal to

$$\partial p/\partial x - y \partial u/\partial y + k \partial v/\partial y.)$$

No other differential condition is independent of (19) when taken together with (16). However, it is necessary that $v = -\delta$ at $(k, 0)$, and hence that (if the integral be taken along the shock from the wall to where it becomes straight)

$$\int \frac{\partial v}{\partial y} dy = \int \frac{B}{y} dp = \delta. \quad (20)$$

When p has been obtained, u and v will be deducible from the second and third of equations (8), which specify their rate of change along radii vectores, starting from

their values on the shock, which are deducible from those of p , and their values on the unit circle, which are known to be zero (except on part of it, in the supersonic case, when they are given by the linearized theory of the Prandtl-Meyer expansion). Since $v = -\delta$ at points of the wall on *both* sides of the origin (by conditions (17), (18) and (20)) and is constant on the two radii vectores with $y = 0$, this boundary condition must be satisfied all along the wall. At $x = y = 0$ both u and v (but not p) are discontinuous, being different on the different radii vectores. But this implies no physical difficulty, since the acceleration of a fluid element,

$$-t^{-1}(x\partial u/\partial x + y\partial u/\partial y, x\partial v/\partial x + y\partial v/\partial y),$$

is finite. This is because the motion of the element in the reduced plane (x, y) is proportional to the vector $-(x, y)$. More difficult is the fact that if (as will appear for all values of M) $\partial p/\partial x \neq 0$ at $x = y = 0$, u must be logarithmically infinite at this point, by (8) (though the acceleration will still be finite); this conflicts with the assumption of small perturbations. However, such a logarithmic infinity occurs in some other 'cone-field' problems, and also in simpler problems treated by the linearized theory where the solution is known to be correct elsewhere and to possess a finite peak corresponding to the logarithmic infinity. The pressure will be found to be bounded (except at the corner in the subsonic case, which is also well-authenticated in other problems), and this is a strong indication of correct approximation of the truth.

The boundary conditions appropriate to equation (16) for p are now complete in the domain $x^2 + y^2 < 1$, $y > 0$, $x > k$; in this domain it is elliptic, and Busemann showed how it can be transformed into Laplace's equation.

4. BUSEMANN'S TRANSFORMATION

In polar co-ordinates with $x = r \cos \theta$, $y = r \sin \theta$, equation (16) becomes

$$\frac{\partial^2 p}{\partial r^2} + \frac{1}{r} \frac{\partial p}{\partial r} + \frac{1}{r^2} \frac{\partial^2 p}{\partial \theta^2} = \left(r \frac{\partial}{\partial r} + 1\right) \left(r \frac{\partial p}{\partial r}\right), \quad (21)$$

and with $\rho = [1 - (1 - r^2)^{\frac{1}{2}}]/r$ this becomes Laplace's equation

$$\frac{\partial^2 p}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial p}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2 p}{\partial \theta^2} = 0 \quad (22)$$

in (ρ, θ) as polar co-ordinates; while the circle $r = 1$ becomes the circle $\rho = 1$, since ρ increases from 0 to 1 monotonically as r does so. Further, $r = 2\rho/(1 + \rho^2)$, so that the line $x = k$ becomes an arc of the circle $2\rho \cos \theta = k(1 + \rho^2)$, cutting $\rho = 1$ orthogonally at $\cos \theta = k$.

The shock boundary condition (19) is transformed into the new co-ordinates by use of the relation

$$\frac{\partial p/\partial x}{\partial p/\partial y} = \frac{(\partial p/\partial r) \cot \theta + \partial p/r \partial \theta}{\partial p/\partial r + (\partial p/r \partial \theta) \cot \theta} = \frac{(r\rho^{-1} dp/dr) (\partial p/\partial \rho) \cot \theta + \partial p/\rho \partial \theta}{(r\rho^{-1} dp/dr) (\partial p/\partial \rho) + (\partial p/\rho \partial \theta) \cot \theta}. \quad (23)$$

But if dn and ds are elements normal and tangential to the circular arc

$$2\rho \cos \theta = k(1 + \rho^2),$$

respectively towards its centre and away from the line $\theta = 0$, and ϕ is the acute angle made by the arc with the radius vector, then

$$\frac{\partial p / \partial \rho}{\partial p / \partial \theta} = \frac{(\partial p / \partial s) \cot \phi + \partial p / \partial n}{\partial p / \partial s - (\partial p / \partial n) \cot \phi}. \quad (24)$$

Hence

$$\frac{\partial p / \partial x}{\partial p / \partial y} = \frac{[(r\rho^{-1} d\rho/dr) \cot \phi \cot \theta - 1] (\partial p / \partial s) + [(r\rho^{-1} d\rho/dr) \cot \theta + \cot \phi] (\partial p / \partial n)}{[(r\rho^{-1} d\rho/dr) \cot \phi + \cot \theta] (\partial p / \partial s) + [(r\rho^{-1} d\rho/dr) - \cot \phi \cot \theta] (\partial p / \partial n)}. \quad (25)$$

But $r\rho^{-1} d\rho/dr = (1 + \rho^2)(1 - \rho^2)$, and on the arc this is $\cos \theta (\cos^2 \theta - k^2)^{-\frac{1}{2}}$. Also the geometry of the triangle formed by the origin, a point (ρ, θ) on the arc, and the point $(k^{-1}, 0)$ which is the centre of the arc, shows that $\cos \phi = (1 - k^2)^{-\frac{1}{2}} \sin \theta$; hence $r\rho^{-1} d\rho/dr = \cot \phi \cot \theta$, and the right-hand side of (25) simplifies to

$$\sin^2 \phi \tan^2 \theta \left[\cot^2 \phi \cot^2 \theta - 1 + \operatorname{cosec}^2 \theta \cot \phi \frac{\partial p / \partial n}{\partial p / \partial s} \right]. \quad (26)$$

The left-hand side being given by (19), (25) now becomes

$$\frac{(A + k)y - Bky^{-1}}{1 - k^2} = \frac{k^2}{1 - k^2} \tan \theta + \frac{(1 - k^2 \sec^2 \theta)^{\frac{1}{2}}}{1 - k^2} \frac{\partial p / \partial n}{\partial p / \partial s}, \quad (27)$$

and (with y replaced by its value $k \tan \theta$ on $x = k$) the condition satisfied by p on the arc $2\rho \cos \theta = k(1 + \rho^2)$ becomes

$$\frac{\partial p / \partial n}{\partial p / \partial s} = \frac{Ak \tan \theta - B \cot \theta}{(1 - k^2 \sec^2 \theta)^{\frac{1}{2}}}. \quad (28)$$

The other boundary conditions are unaltered, except for the discontinuity condition (18) at the corner in the subsonic case. Putting $\xi = \rho \cos \theta$, $\eta = \rho \sin \theta$, the corner becomes the point $\xi = -(1 - (1 - M_1^2)^{\frac{1}{2}})/M_1$, $\eta = 0$, and the condition holding there is that

$$\lim_{\eta \rightarrow 0} \int \left(\frac{\rho}{r} \frac{\partial p}{\partial \eta} \right) \left(\frac{dr}{d\rho} d\xi \right) = M_1 \delta, \quad (29)$$

i.e.
$$\lim_{\eta \rightarrow 0} \int \frac{\partial p}{\partial \eta} d\xi = [r\rho^{-1} d\rho/dr]_{r=M_1} M_1 \delta = \frac{M_1 \delta}{(1 - M_1^2)^{\frac{1}{2}}}. \quad (30)$$

5. SOLUTION OF THE POTENTIAL PROBLEM

Now p is given as a harmonic function satisfying certain boundary conditions in a curvilinear triangle ABC with AB and BC circular arcs and AC a straight segment, and all its angles right angles. To solve this potential problem conformal transformation into a simpler domain is necessary.

Now geometrical inversion of this curvilinear triangle with respect to B will produce a triangle $A'B'C'$ with B' at infinity, $A'B'$ and $B'C'$ straight segments and $A'C'$ a circular arc, and all its angles (which are unaltered by inversion) right angles. But if $\zeta = \rho e^{i\theta}$, the transformation

$$Z = (k + ik') \left(i - \frac{2k'}{\zeta - (k + ik')} \right), \quad (31)$$

where B is the point $\zeta = k + ik'$, i.e. $k' = (1 - k^2)^{\frac{1}{2}}$, performs this inversion, together with reflexion, translation and rotation which do not affect the qualitative statement. Further, when ζ is on the arc $2\rho \cos \theta = k(1 + \rho^2)$,

$$Z = \frac{(\cos^2 \theta - k^2)^{\frac{1}{2}}}{k' \cos \theta - k \sin \theta}, \quad (32)$$

which is purely real and increases from 1 to ∞ as θ increases from 0 to

$$\tan^{-1}(k'/k) = \cos^{-1} k.$$

Hence the triangle $A'B'C'$ must consist of that part of the first quadrant which is outside the unit circle.

Equation (31) can be solved for θ as

$$\tan \theta = \frac{k'(Z^2 - 1)}{k(Z^2 + 1)}, \quad (33)$$

hence in the Z -plane the boundary condition (27) becomes

$$\frac{\partial p / \partial n}{\partial p / \partial s} = \left\{ \frac{Ak'(Z^2 - 1)}{Z^2 + 1} - \frac{Bk(Z^2 + 1)}{k'(Z^2 - 1)} \right\} / \left(\frac{2k'Z}{Z^2 + 1} \right). \quad (34)$$

The left-hand side is unaltered (since in conformal transformation the local deformation of plane elements is a pure dilatation) provided that dn is still an outward normal to the domain and that ds is in the positive tangential direction.

Lastly, with

$$z_1 = \frac{1}{2}(Z^2 + Z^{-2}), \quad Z^2 = z_1 + (z_1^2 - 1)^{\frac{1}{2}}, \quad (35)$$

the z_1 -domain corresponding to the triangle $A'B'C'$ is the upper half-plane. The shock, to which the part of the real axis with $Z > 1$ corresponded, becomes the part of the real axis with $z_1 > 1$. The wall, to which the first quadrant of the circle $|Z| = 1$ corresponded, becomes the part of the real axis with $-1 < z_1 < 1$. The third boundary $r = 1$, or $\rho = 1$, to which the part of the imaginary axis with $\Im Z > 1$ corresponded, becomes the part of the real axis with $z_1 < -1$.

The shock boundary condition (34) becomes (with $z_1 = x_1 + iy_1$) the condition that

$$\frac{\partial p / \partial y_1}{\partial p / \partial x_1} = \left[B \frac{k}{1 - k^2} (x_1 + 1) - A(x_1 - 1) \right] [2(x_1 - 1)]^{-\frac{1}{2}} \quad (36)$$

on $x_1 > 1$, $y_1 = 0$.

The wall boundary condition is that $\partial p / \partial y_1 = 0$ when $-1 < x_1 < 1$, $y_1 = 0$. The discontinuity condition (29) becomes

$$\lim_{y_1 \rightarrow 0} \int \frac{\partial p}{\partial y_1} dx_1 = \frac{M_1 \delta}{(1 - M_1^2)^{\frac{1}{2}}}, \quad (37)$$

and holds at the point corresponding to $\zeta = -(1 - (1 - M_1^2)^{\frac{1}{2}}) / M_1$, which is

$$z_1 = -\frac{(M_1 + k)^2 + (M_1^2 - 1)(1 - k^2)}{(M_1 k + 1)^2} = x_0 > -1. \quad (38)$$

The condition on the third boundary can be written $\partial p / \partial x_1 = 0$ when $x_1 < -1$, $y_1 = 0$. But in the supersonic case this must be supplemented with

$$\lim_{y_1 \rightarrow 0} \int \frac{\partial p}{\partial x_1} dx_1 = -\frac{M_1 \delta}{(M_1^2 - 1)^{\frac{1}{2}}}, \quad (39)$$

by (17), an equation which holds at the point in the z_1 -plane corresponding to $\zeta = -M_1^{-1} + i(1 - M_1^{-2})^{\frac{1}{2}}$; but this point is found to have exactly the same algebraic expression as the x_0 of (38), only in the supersonic case $x_0 < -1$.

The solution is effected by the introduction of a function

$$w(z_1) = \partial p / \partial y_1 + i \partial p / \partial x_1,$$

which is regular throughout the upper half-plane since p is harmonic. In terms of w , the discontinuity conditions (38) and (39) can be expressed by saying that, near $z_1 = x_0$,

$$w \sim \frac{M_1 \delta / \pi (M_1^2 - 1)^{\frac{1}{2}}}{z_1 - x_0} \text{ (supersonic), } w \sim \frac{i M_1 \delta / \pi (1 - M_1^2)^{\frac{1}{2}}}{z_1 - x_0} \text{ (subsonic).} \quad (40)$$

Further, w is real on $x_1 < -1$, $y_1 = 0$ and purely imaginary on $-1 < x_1 < 1$, $y_1 = 0$. Lastly, on $x_1 > 1$, $y_1 = 0$,

$$\arg w(z_1) = \arctan \frac{[2(x_1 - 1)]^{\frac{1}{2}}}{Bk(1 - k^2)^{-1}(x_1 + 1) - A(x_1 - 1)} = \arctan \frac{(x_1 - 1)^{\frac{1}{2}}}{\alpha} + \arctan \frac{(x_1 - 1)^{\frac{1}{2}}}{\beta}, \quad (41)$$

$$\text{where } \frac{\alpha \beta \sqrt{2}}{\alpha + \beta} = \frac{2Bk}{1 - k^2}, \quad \frac{\sqrt{2}}{\alpha + \beta} = A - \frac{Bk}{1 - k^2}, \quad (42)$$

so that, by (9) and (15),

$$\alpha \beta = 2M^2, \quad \alpha + \beta = 2M^2 k \sqrt{2}, \quad (43)$$

whence (since $Mk > 1$) both α and β are positive.

A function satisfying equation (41) is

$$[\alpha - i(z_1 - 1)^{\frac{1}{2}}]^{-1} [\beta - i(z_1 - 1)^{\frac{1}{2}}]^{-1}, \quad (44)$$

where $(z_1 - 1)^{\frac{1}{2}}$ has non-negative imaginary part in the upper half-plane. This function is regular throughout the domain, and even on the boundary for $z_1 \neq 1$; it is real and positive for $x_1 < 1$, $y_1 = 0$. The condition (41) will still be satisfied if (44) be multiplied by any function real for $x_1 > 1$, $y_1 = 0$. The functions chosen are, first, $(z_1^2 - 1)^{-\frac{1}{2}}$ (the branch positive for $x_1 > 1$, $y = 0$; it is purely imaginary for $-1 < x_1 < 1$, $y_1 = 0$); secondly, $-C\delta(z_1 - x_0)^{-1}$, where C is a constant to be determined by conditions (40); and thirdly, $1 - D(z_1 - x_0)$, where D is to be determined by condition (20). The combined expression

$$w(z_1) = \frac{C\delta[D(z_1 - x_0) - 1]}{(z_1^2 - 1)^{\frac{1}{2}}(z_1 - x_0)[\alpha - i(z_1 - 1)^{\frac{1}{2}}][\beta - i(z_1 - 1)^{\frac{1}{2}}]} \quad (45)$$

is the unique function satisfying all the boundary conditions and integrable at every point including infinity but excluding x_0 .

The constant C , by (40), satisfies

$$C = \frac{[\alpha + (1 - x_0)^{\frac{1}{2}}][\beta + (1 - x_0)^{\frac{1}{2}}] |1 - x_0^2|^{\frac{1}{2}} M_1}{\pi |1 - M_1^2|^{\frac{1}{2}}} > 0. \quad (46)$$

It has a finite non-zero limit for $M_1 = 1$, $x_0 = -1$. Since, by (33) and (35), on $x_1 > 1$, $y_1 = 0$, y is given by

$$y = k' \left(\frac{x_1 - 1}{x_1 + 1} \right)^{\frac{1}{2}}, \quad (47)$$

equation (20) becomes

$$\begin{aligned}\delta &= \int_1^\infty \frac{B}{y} \frac{\partial p}{\partial x_1} dx_1 = \int_1^\infty \frac{B}{k'} \left(\frac{x_1+1}{x_1-1} \right)^{\frac{1}{2}} \frac{C\delta[D(x_1-x_0)-1](\alpha+\beta)(x_1-1)^{\frac{1}{2}}}{(x_1^2-1)^{\frac{1}{2}}(x_1-x_0)(\alpha^2+x_1-1)(\beta^2+x_1-1)} dx_1 \\ &= \frac{BC\delta(\alpha+\beta)}{k'} \int_0^\infty \frac{D(x+\gamma^2)-1}{x^{\frac{1}{2}}(x+\alpha^2)(x+\beta^2)(x+\gamma^2)} dx \\ &= \frac{BC\delta(\alpha+\beta)}{k'} \pi \frac{(D\gamma^2-1)(\alpha+\beta+\gamma)/\alpha\beta\gamma+D}{(\alpha+\beta)(\beta+\gamma)(\gamma+\alpha)},\end{aligned}\quad (48)$$

where

$$\gamma^2 = 1 - x_0 = 2(M_1 + k)^2 / (M_1 k + 1)^2. \quad (49)$$

Since by (46)

$$\begin{aligned}\frac{C}{(\beta+\gamma)(\gamma+\alpha)} &= \frac{|1-x_0^2|^{\frac{1}{2}} M_1}{\pi |1-M_1^2|^{\frac{1}{2}}} = \frac{2(M_1+k)^2 \frac{1}{2} (1-M_1^2)(1-k^2)^{\frac{1}{2}}}{(M_1 k + 1)^2 (M_1 k + 1)^2} \left| \frac{M_1}{\pi |1-M_1^2|^{\frac{1}{2}}} \right| \\ &= \frac{2M_1 k' (M_1 + k)}{\pi (M_1 k + 1)^2},\end{aligned}\quad (50)$$

the condition determining D becomes

$$\frac{(M_1 k + 1)^2}{2BM_1(M_1 + k)} = D \frac{(\gamma + \alpha)(\gamma + \beta)}{\alpha\beta} = \frac{\alpha + \beta + \gamma}{\alpha\beta\gamma}, \quad (51)$$

whence it is easily calculated using (43).

The solution is now complete, and only back-transformation and computation are necessary for the deduction of physical results. These have been computed, and are set out in §§ 6-8, for the following values of M : 1.36277, 1.64751, 2.06809, 2.95200, ∞ . (The limit as $M \rightarrow 1$ is discussed in § 9.) The corresponding shock pressure-ratios p_1/p_0 , Mach numbers behind the shock M_1 , and positions of the shock (relative to the radius of the circle of propagation) k are given in table 1; these explain the choice of values.

TABLE 1

M	1	1.36277	1.64751	2.06809	2.95200	∞
p_1/p_0	1	2	3	4.82315	10	∞
M_1	0	0.47245	0.72739	1	1.34463	1.88982
k	1	0.75593	0.65465	0.56619	0.47809	0.37796

As $M \rightarrow \infty$, by (43), $\alpha \sim 2M^2 k \sqrt{2}$ and $\beta \rightarrow 1/k\sqrt{2}$. Hence by (45) and (50) the limiting expression for $w(z_1)$ is

$$\frac{2M_1 k' (M_1 + k) (\beta + \gamma)}{\pi (M_1 k + 1)^2} \frac{\delta[D(z_1 - x_0) - 1]}{(z_1^2 - 1)^{\frac{1}{2}} (z_1 - x_0) [\beta - i(z_1 - 1)^{\frac{1}{2}}]}, \quad (52)$$

which is finite and in no way more singular than (45). Thus the pressure and velocity fields reduced in scale (proportionately to the pressure and velocity behind the straight part of the shock) have a finite limiting shape and distribution as $M \rightarrow \infty$; these are treated below on the same footing as those for the finite values of M considered.

6. THE SHAPE OF THE DIFFRACTED SHOCK

At a point (k, y) of the shock, or in the (x_1, y_1) plane $(x_1, 0)$, the shock curvature (taken positive when the shock is convex to the still air) is

$$\begin{aligned}\kappa = -f''(y) &= \frac{\partial v}{\partial y} = \frac{B}{y} \frac{\partial p}{\partial y} = \frac{B}{y} \frac{dx_1}{dy} \frac{\partial p}{\partial x_1} = \frac{B(x_1+1)^2}{1-k^2} \frac{\partial p}{\partial x_1} \\ &= \frac{BC\delta(\alpha+\beta)(x_1+1)^{\frac{3}{2}}[D(x_1-x_0)-1]}{(1-k^2)(x_1-x_0)(\alpha^2+x_1-1)(\beta^2+x_1-1)},\end{aligned}\quad (53)$$

by (47) and (45). κ/δ is graphed against y/k' (which runs from 0 to 1 on the curved part of the shock), for the four finite values of M listed above, in figure 1, where the limit as $M \rightarrow \infty$, which is

$$\frac{\kappa}{\delta} = \frac{2M_1 k' (M_1 + k) (\beta + \gamma)}{\pi (M_1 k + 1)^2} \frac{B(x_1+1)^{\frac{3}{2}} [D(x_1-x_0)-1]}{(1-k^2)(x_1-x_0)(\beta^2+x_1-1)}, \quad (54)$$

is also shown. While (53) is zero at $x_1 = \infty$ ($y = k'$), (54) is infinite; the figure shows that this infinity is the limit of a very steep peak, which occurs just before $y = k'$ for $M < \infty$. The infinity is only one of curvature, and leads to no great peculiarity of shape of the shocks themselves.

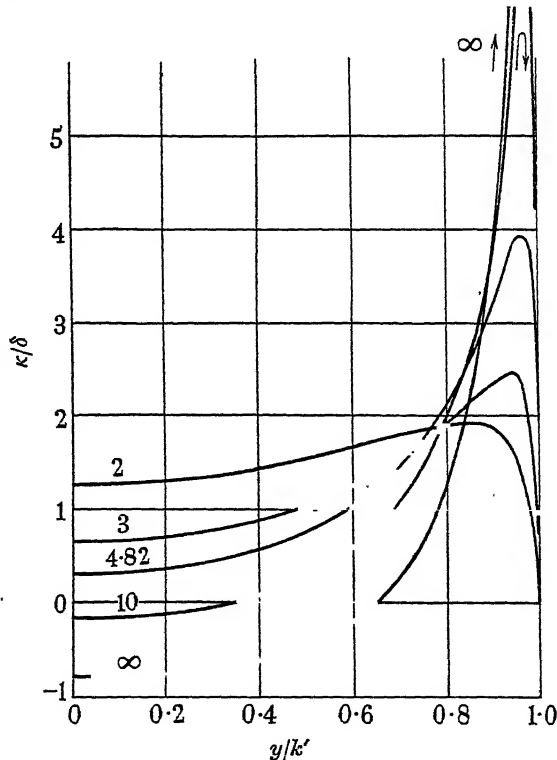


FIGURE 1. Curvature of diffracted shock. The numbers on the curves are the values of p_1/p_0 .

It is observed that, for the larger values of M (or of p_1/p_0), κ is negative for small y ; so that the shock is concave to the still air near the wall, changing to convex farther out through a point of inflexion. This is due to the term in square brackets in (53)

and (54); the point of inflexion is at $x_1 = x_0 + D^{-1}$, and appears if $x_0 + D^{-1} > 1$. The transition occurs when $D = \gamma^{-2}$, hence by (51) and (49) when

$$1 = \gamma^2 \frac{(M_1 k + 1)^2}{2BM_1(M_1 + k)} = \frac{M_1 + k}{BM_1}; \quad (55)$$

whence, by (3), (9) and (15), $M = 2.53111$ and $p_1 = 7.30760p_0$. Hence it is for shock pressure-ratios > 7.31 that a point of inflexion should appear in the diffracted shock. It will be observed that the maximum weakening at the wall for an entirely convex shock, which would be obtained by a function $\kappa(y)$ always non-negative but entirely concentrated near $y = k'$ (and still satisfying the condition $\int_0^1 (\kappa/\delta) d(y/k') = 1/k'$, which is a check on each of the curves in figure 1), would be obtained by drawing the diffracted shock as a straight line normal to the wall: but that greater weakening is possible once a point of inflexion is allowed.

The shapes of the shocks, and of the regions behind them which are disturbed from uniform flow, for $\delta = 0.1$ radian, are shown, together with the wall pressure distributions, in the five cases considered, in figures 2 to 6, whose scales are chosen to make the distance between the corner and the shock the same in each; the thick line represents the shock.

7. PRESSURE DISTRIBUTION ALONG THE WALL

At a point $(x, 0)$ of the wall ($-1 < x < k$), the x_1 co-ordinate is

$$x_1 = 1 - 2 \frac{(k-x)^2}{(1-kx)^2}, \quad (56)$$

and satisfies $-1 < x_1 < 1$. Now in this region, by (45),

$$\frac{\partial p}{\partial x_1} = - \frac{C\delta[D(x_1 - x_0) - 1]}{(1 - x_1^2)^{\frac{1}{2}}(x_1 - x_0)[\alpha + (1 - x_0)^{\frac{1}{2}}][\beta + (1 - x_0)^{\frac{1}{2}}]} \quad (57)$$

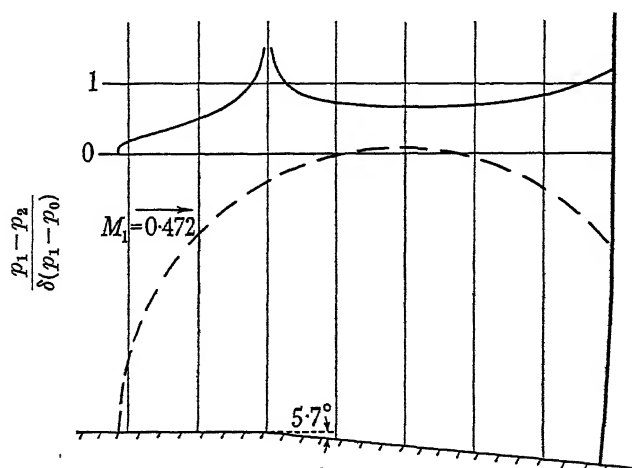
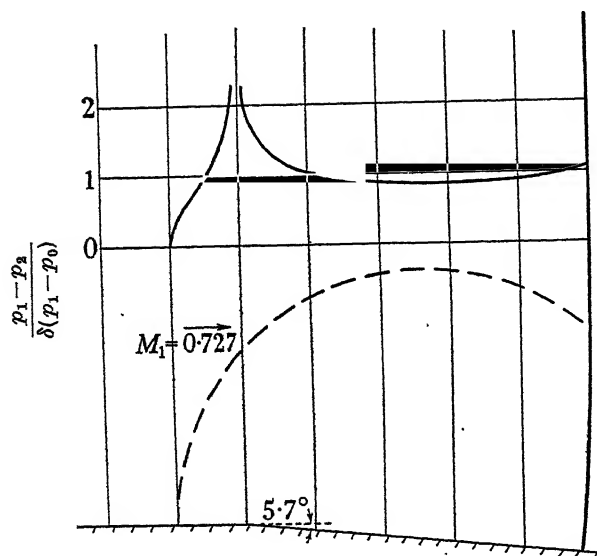
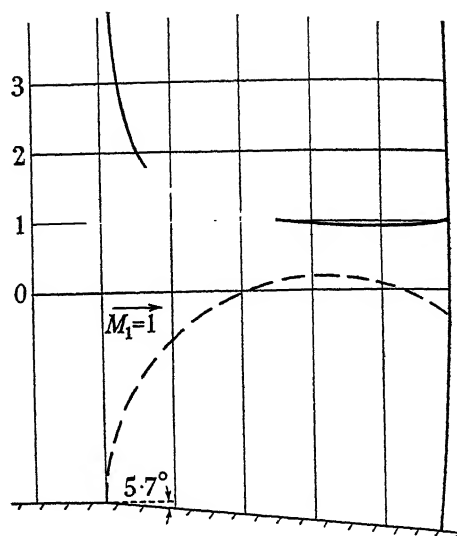
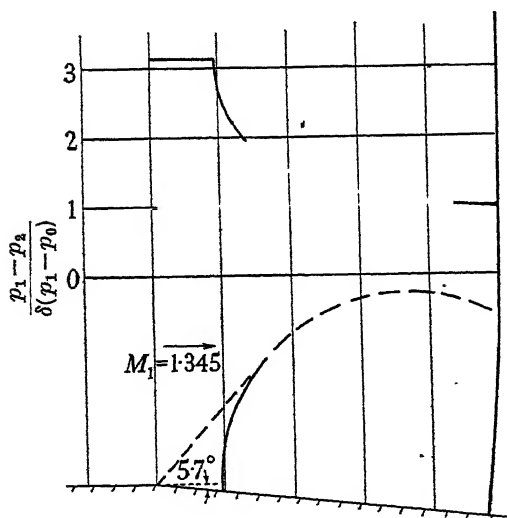
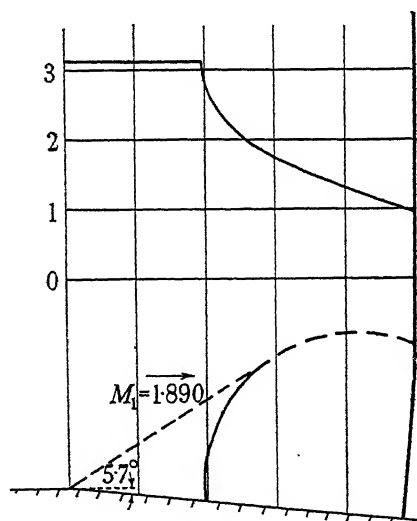


FIGURE 2. Wall pressure distribution and shape of disturbed region ($\delta = 0.1$ radian, $p_1/p_0 = 2$).

FIGURE 3. ($\delta = 0.1$ radian, $p_1/p_0 = 3$.)FIGURE 4. ($\delta = 0.1$ radian, $p_1/p_0 = 4.82$.)

FIGURES 3, 4. Wall pressure distribution and shape of disturbed region.

FIGURE 5. ($\delta = 0.1$ radian, $p_1/p_0 = 10$.)FIGURE 6. ($\delta = 0.1$ radian, $p_1/p_0 = \infty$.)

FIGURES 5, 6. Wall pressure distribution and shape of disturbed region.

(with a similar expression corresponding to (52) for $M \rightarrow \infty$), and while p could theoretically be integrated from (57) as an elementary function it is a simpler matter in computation to use numerical integration. The integration is best carried out over equal intervals of the angle $\cos^{-1} x_1$, a device which eliminates the singularities at $x_1 = \pm 1$. When $M_1 \leq 1$, so that $-1 \leq x_0 < 1$, the known singularity of (57) at $x_1 = x_0$ must be subtracted before integration, and the integrated singular part added to the result.

The quantity chosen to be graphed against x in figures 2 to 6 is the relative deficiency of pressure at the wall, divided by the angle δ in radians, that is,

$$\frac{p_1 - p_2}{\delta(p_1 - p_0)} = \frac{a_1 q_1 \rho_1}{p_1 - p_0} \left(-\frac{p}{\delta} \right) = -\frac{p}{k\delta}. \quad (58)$$

The qualitative results which appear could have been simply deduced from the form of (57): that in the subsonic case the deficiency rises from zero at the boundary of the disturbed region to a logarithmic infinity at the corner, then falls to a minimum and rises again to its final value at the shock: that in the sonic case ($M_1 = 1$) the deficiency falls from its algebraically infinite value (to which at the corner it has suddenly jumped from zero) to a minimum and then rises: that in the supersonic case it has a Prandtl-Meyer discontinuity at the corner from which, in $-1 < x < k$, it falls and then slightly rises again when $p_1/p_0 < 7.31$, but falls monotonically when $p_1/p_0 > 7.31$. The logarithmic infinity in the subsonic case corresponds to a rapid flow round the corner. In the sonic case the true $(p_1 - p_2)/(p_1 - p_0)$ immediately behind the corner is probably equal to the Prandtl-Meyer value with sonic initial velocity, which is of order $\delta^{\frac{1}{2}}$ for small δ , so that it is not surprising that the present theory gives $\lim (p_1 - p_2)/\delta(p_1 - p_0) = \infty$ at this point.

Quantitatively it is seen that the scale of even the *relative* pressure deficiency $(p_2 - p_1)/(p_1 - p_0)$ increases with the shock pressure-ratio p_1/p_0 . All the pressure distributions given have been checked satisfactorily by calculating the pressure at the junction of wall and shock independently by the integration of $y f''(y)$ as obtained in § 6, using equation (13).

8. NATURE OF THE THIRD BOUNDARY FOR CONVEX AND CONCAVE CORNERS

From equation (45) it is possible to calculate $\partial p/\partial y_1$ on $x_1 < -1$, $y_1 = 0$; and, in Busemann's (ρ, θ) plane, the value of the pressure-derivative $\partial p/\partial n$, along the *inward* normal on that part of the boundary with $\rho = 1$, is deduced therefrom by multiplying by a factor $|dz_1/d\zeta|$ which varies but remains positive. Thus, when $\delta > 0$,

$$\frac{\partial p}{\partial n} = \left| \frac{dz_1}{d\zeta} \right| \frac{C\delta[D(x_1 - x_0) - 1]}{-(x_1^2 - 1)^{\frac{1}{2}}(x_1 - x_0)[\alpha + (1 - x_1)^{\frac{1}{2}}][\beta + (1 - x_1)^{\frac{1}{2}}]}, \quad (59)$$

which is negative when $x_1 < x_0$, but is positive when $x_0 < x_1 < -1$ (provided also $x_1 < x_0 + D^{-1}$; but actually $x_0 + D^{-1} \geq -1$ for all M).

But in the original plane (r, θ) , $(\partial p/\partial r)_{r=1}$ is infinite, since it is obtained from (59) by multiplication by $-[d\rho/dr]_{r=1} = \infty$. The actual behaviour near $r = 1$ is deduced from the asymptotic equality $1 - \rho \sim [2(1 - r)]^{\frac{1}{2}}$ as

$$p - (p)_{r=1} \sim \frac{\partial p}{\partial n}(1 - \rho) \sim \frac{\partial p}{\partial n}[2(1 - r)]^{\frac{1}{2}}. \quad (60)$$

This behaviour is probably not what really occurs but is a singularity which is the only way that the linearized theory knows of describing some more complicated phenomenon, such as a shock or a rapid expansion. The same behaviour of the solution occurs in the cone-field theory of supersonic steady flow, referred to in § 3.

Experience in the latter field points to the tentative conclusion that the true phenomenon is a shock when $\partial p/\partial n > 0$ and an expansion (rapid but not discontinuous) when $\partial p/\partial n < 0$.* In the latter case the boundary must be the circle exactly, since it must be a characteristic: but in the former the shock must be slightly further from the origin than the circle, in order that its own motion shall be supersonic.

On this assumption, and taking $\delta > 0$, the whole circle represents an expansion in the subsonic case: it is accordingly dotted in figures 2, 3 and 4. But in the supersonic case $x_0 < -1$, and the circle is an expansion for $x_1 < x_0$, i.e. for points to the right of the point of contact of the tangent from the corner; and this part, together with the tangent representing a Prandtl-Meyer expansion, is shown dotted in figures 5 and 6; but the remainder of the circle, shown plain, has $\partial p/\partial n > 0$ and probably corresponds to a weak shock.

But when $\delta < 0$, so that the corner is concave to the flow, the sign of $\partial p/\partial n$ is changed throughout and the whole circle corresponds to a shock in the subsonic case, which in the supersonic case is replaced by part of the circle plus the tangent from the corner. The shock should be drawn slightly away from the circle, except where it joins the main shock ($x = k, x_1 = \infty$), when by (59) $\partial p/\partial n = 0$. (The exact position of the straight oblique shock in the supersonic case is known from steady flow theory.) The shock pattern is sketched in two cases ($p_1/p_0 = 3$ and 10) in figure 7. The type of three shock intersection that occurs is similar to the well-known experimental phenomenon of 'Mach reflexion'.

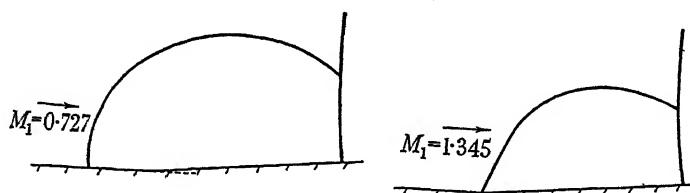


FIGURE 7. Corner concave to the flow: three shock intersections ($\delta = -0.1$ radian, $p_1/p_0 = 3, 10$).

9. LIMITING SOLUTION AS $M \rightarrow 1$: CONNEXION WITH THE ACOUSTIC THEORY

As $M \rightarrow 1$ the origin of (x, y) co-ordinates approaches the corner, which in the limit becomes the centre of the circle of disturbance, to which the undisturbed portion of the shock is a tangent. In the acoustic theory of diffraction round a corner (Sommerfeld 1901; Friedlander 1946), with δ not small, the diffracted wave front is similarly a circular arc with the corner as centre and the undisturbed wave front as tangent. In this section the wall pressure distribution, given by the present theory (which has treated the limit of the problem as $\delta \rightarrow 0$) when $M \rightarrow 1$, is compared with that given by the acoustic theory (which lets $M \rightarrow 1$ at the start) when $\delta \rightarrow 0$. The inversion of this double limit is found to be valid, but not uniformly so near the wall-shock junction.

Let $M = 1 + \epsilon$: then from (9) $1 - k \sim \epsilon$ as $\epsilon \rightarrow 0$. By (56), for fixed $x < 1$ as $\epsilon \rightarrow 0$,

$$x_1 + 1 \sim 4\epsilon \frac{1+x}{1-x}. \quad (61)$$

* The author will shortly publish a proof that this is so in supersonic steady flow.

But α, β, γ all tend to $\sqrt{2}$ as $\epsilon \rightarrow 0$, so by (50) and (51)

$$C \sim 16M_1 k'/\pi \quad \text{and} \quad D \sim \frac{1}{8}(BM_1)^{-1} \sim \frac{3}{40}\epsilon^{-2}.$$

Hence in (57) $(x_1 - x_0)^{-1}$ can be neglected (in the limit) in comparison with D , if $x \neq 0$; and

$$\begin{aligned} \frac{\partial p}{\partial x} &\sim \frac{dx_1}{dx} \frac{\delta(-16M_1 k'/\pi) \frac{1}{8}(BM_1)^{-1}}{[8\epsilon(1+x)/(1-x)]^{\frac{1}{2}} 8} \sim \frac{8\epsilon}{(1-x)^2} \frac{\delta[-16\sqrt{(2\epsilon)/\pi}] (\frac{1}{8}\epsilon^{-1})}{[8\epsilon(1+x)/(1-x)]^{\frac{1}{2}} 8} \\ &= -\frac{\delta/\pi}{(1-x)^{\frac{1}{2}}(1+x)^{\frac{1}{2}}}. \end{aligned} \quad (62)$$

The reasoning is not uniformly correct near $x = 1$, since (61) fails here. The value of p at the wall-shock junction (deduced from (57)) as $\epsilon \rightarrow 0$ is in fact

$$\begin{aligned} -CD\delta \int_{-1}^1 \frac{dx_1}{(1-x_1^2)^{\frac{1}{2}} [\sqrt{2+(1-x_1)^2}]^2} \\ = -\frac{2\sqrt{2} \cdot \delta}{\pi \sqrt{\epsilon}} \int_0^1 \frac{dy}{(1-y^2)^{\frac{1}{2}} (1+y)^2} = -\frac{4\sqrt{2}}{3\pi} \frac{\delta}{\sqrt{\epsilon}}. \end{aligned} \quad (63)$$

(But this does not mean an infinite pressure as $\epsilon \rightarrow 0$ since the actual pressure differs from p_1 by $a_1 q_1 \rho_1 p \sim \frac{7}{3}\epsilon p_0 p$).

On the acoustic theory there is no discontinuity at the refracted part of the shock (Sommerfeld 1901). Probably this is because in fact the discontinuity is $o(\epsilon)$ as $\epsilon \rightarrow 0$. Thus, in the notation of the present paper, while $p = 0$ for $r = 1$, $0 < \theta < \pi$, the equation holding for $-\delta < \theta < 0$ is

$$p = \frac{p_0 - p_1}{a_1 q_1 \rho_1} = -\frac{\rho_0 U}{a_1 \rho_1} \sim -1, \quad (64)$$

where equation (1) has been used. Now p satisfies (16) as before, and hence also (22). The ζ -plane is a sector of angle $\pi + \delta$. The transformations

$$(\zeta e^{\pi i \delta})^{\pi/(\pi+\delta)} = Z_1, \quad -\frac{1}{2}(Z_1 + Z_1^{-1}) = z_2, \quad (65)$$

transform it into the upper half z_2 -plane; in which the boundary conditions are $\partial p/\partial y_2 = 0$ for $y_2 = 0$, $|x_2| > 1$; $p = 0$ for $y_2 = 0$, $-\cos[\pi^2/(\pi+\delta)] < x_2 < 1$; $p = -1$ for $y_2 = 0$, $-1 < x_2 < -\cos[\pi^2/(\pi+\delta)]$.

The solution with p bounded is given by

$$\frac{\partial p}{\partial y_2} + i \frac{\partial p}{\partial x_2} = \frac{-\pi^{-1} \sin[\pi^2/(\pi+\delta)]}{(z_2 + \cos[\pi^2/(\pi+\delta)])(1-z_2^2)^{\frac{1}{2}}}, \quad (66)$$

which yields in $|x_2| > 1$, $y_2 = 0$

$$\frac{\partial p}{\partial x_2} = \frac{-\pi^{-1} \sin[\pi^2/(\pi+\delta)] \operatorname{sgn} x_2}{(x_2 + \cos[\pi^2/(\pi+\delta)])(x_2^2 - 1)^{\frac{1}{2}}} \sim \frac{-\pi^{-1} \delta \operatorname{sgn} x_2}{(x_2 - 1)(x_2^2 - 1)^{\frac{1}{2}}}, \quad (67)$$

as $\delta \rightarrow 0$. But

$$x_2 \sim -\frac{1}{2} \left(\xi + \frac{1}{\xi} \right) = -\frac{1}{2} \left[\frac{1 - (1-x^2)^{\frac{1}{2}}}{x} + \frac{1 + (1-x^2)^{\frac{1}{2}}}{x} \right] = -\frac{1}{x}, \quad (68)$$

so that

$$\frac{\partial p}{\partial x} \sim \frac{1}{x^2} \frac{-\pi^{-1} \delta \operatorname{sgn} x}{(x^{-1} - 1)(x^{-2} - 1)^{\frac{1}{2}}} = \frac{-\delta/\pi}{(1-x)(1-x^2)^{\frac{1}{2}}}, \quad (69)$$

- Busemann, A. 1943 *Luftfahrtforsch.* 20, 105.
 Friedlander, F. G. 1946 *Proc. Roy. Soc. A*, 186, 322.
 Sommerfeld, A. 1901 *Z. Math. phys.* 46, 11.

Kinetic theory of diffusion in gases and liquids.

II. General kinetic theory of liquids mixtures

By L. M. YANG, *University of Edinburgh*

(Communicated by M. Born, F.R.S.—Received 8 December 1948—

Revised 11 March 1949)

In this paper the classical kinetic theory of simple liquids developed by Born & Green is generalized for liquid mixtures. Though the fundamental equations preserve many of the general features of those relating to simple liquids, emphasis is laid on certain points characteristic of a liquid mixture. In the expansion procedure, improvement has been made in the choice of parameters; the potentials of the external force fields are taken as parameters instead of the external forces in order that force diffusion should follow naturally. The hydrodynamical equations for a liquid mixture follow after a set of relations between the observable quantities and those defined with respect to each component in the mixture have been obtained. These equations when compared with the corresponding ones for gas mixtures reveal the peculiar complicated features of liquid mixtures.

A practical method of solution for the expanded equation of motion is discussed and general expressions for the coefficients of ordinary diffusion and thermal diffusion and also the coefficient of viscosity in a liquid mixture are given in the last section.

1. INTRODUCTION

While the kinetic theory of gases and solids has progressed a long way towards perfection, little has been done on the kinetic theory of liquids owing to the fact that no simple assumptions are available as in the case of gases or solids. There are, however, quite a number of fragmentary theories devised for special purposes as summarized by Frenkel (1946). Recently in a series of papers by Born & Green (1946, 1947 *a, b*), and independently by Kirkwood (1946, 1947), a new attempt has been made towards a rigorous and systematic formulation of the kinetic theory of liquids. The essential feature of this theory is the introduction of multiple distribution functions which make possible the study of molecular clusters. Though a convenient basis has been set up, there remains much to be done to bring the theory nearer to the actual calculation of empirical constants from atomic data. The present paper deals with the case of a liquid mixture, proceeding along much the same line as part III of Born & Green's paper. Fundamental equations of conservation for a molecular cluster are derived from Liouville's theorem in classical phase space. These equations are shown to reduce to the macroscopic hydrodynamical equations with the help of a set of relations between quantities actually observed in a liquid mixture and those defined with reference to the molecules of a particular kind in a mixture. Comparisons are made on several occasions with the example of a gas mixture, thus enabling the peculiar complicated features in a liquid mixture to be seen clearly.

In the expansion procedure, improvement has been made in the choice of physical parameters; the potentials of the molecules in the external force field instead of the external forces themselves are taken as parameters in order that force diffusion may follow naturally from the expanded equation of motion. An exact analogue of the theory of gas mixtures is obtained in the expanded equation of motion except that

the collision integrals in the gas theory are replaced by certain integrals involving the velocity-distribution functions of pairs of molecules, and that there are additional terms involving the potential pressures which are taken to be zero in the gas theory. A practical method of solution is discussed in § 6 which consists of modifying the assumption of binary encounters in the following manner. As one cannot expect to consider interactions of several molecules at a time, the expedient thing to do is to consider the motion of one molecule relative to another in an averaged field due to the presence of the others. This field is taken to be the same as if the fluid were in equilibrium. By this method the pair-distribution functions are expressed in terms of single-distribution functions, and a closed integral equation, which may be regarded as a generalization of Boltzmann's equation, is obtained. No numerical calculation is attempted in the present paper, as this will necessarily be a considerable task. General expressions for the coefficients of diffusion and viscosity obtained from quite general considerations are given in the last section.

2. NOTATION AND DEFINITIONS

Consider an assembly of two different kinds of molecules of total number N occupying a total volume V . Let N_1, N_2 be the numbers of molecules of the first and second kind respectively, and m_1, m_2 their molecular masses. Of these $N (= N_1 + N_2)$ molecules, one can choose any h_1 of the first kind and any h_2 of the second kind for special consideration.

Let the velocity of a molecule of the first kind at the position $\mathbf{x}_1^{(i)}$ be denoted by $\xi_1^{(i)}$. To abbreviate, we shall use \mathbf{x}_1 and ξ_1 to denote all the position and velocity vectors of the h_1 molecules and \mathbf{x}_2 and ξ_2 those of the h_2 molecules. The volume elements $\prod_{i=1}^{h_1} \prod_{j=1}^{h_2} d\mathbf{x}_1^{(i)} d\mathbf{x}_2^{(j)}$ will be denoted by $d\mathbf{x}_{h_1 h_2}$ and $\prod_{i=1}^{h_1} \prod_{j=1}^{h_2} d\xi_1^{(i)} d\xi_2^{(j)}$ by $d\xi_{h_1 h_2}$.

Next, multiple-distribution functions symmetrized with respect to molecules of the same kind will be defined. The distribution function $n_{h_1 h_2} = n_{h_1 h_2}(t, \mathbf{x}_1, \mathbf{x}_2)$ is defined such that $n_{h_1 h_2} d\mathbf{x}_{h_1 h_2}$ is the probability of finding the volume elements $d\mathbf{x}_1^{(1)} \dots d\mathbf{x}_1^{(h_1)}$ occupied by any h_1 molecules of the first kind and the volume elements $d\mathbf{x}_2^{(1)} \dots d\mathbf{x}_2^{(h_2)}$ by any h_2 molecules of the second kind simultaneously at time t . Similarly $f_{h_1 h_2} = f_{h_1 h_2}(t, \mathbf{x}_1, \mathbf{x}_2, \xi_1, \xi_2)$ is defined such that $f_{h_1 h_2} d\mathbf{x}_{h_1 h_2} d\xi_{h_1 h_2}$ is the probability of finding the volume elements $\mathbf{x}_1^{(i)}, d\mathbf{x}_1^{(i)}$ ($i = 1, 2, \dots, h_1$) occupied by any molecules of the first kind with velocities in $\xi_1^{(i)}, d\xi_1^{(i)}$ ($i = 1, 2, \dots, h_1$), and $\mathbf{x}_2^{(j)}, d\mathbf{x}_2^{(j)}$ ($j = 1, 2, \dots, h_2$) occupied by any molecules of the second kind with velocities in $\xi_2^{(j)}, d\xi_2^{(j)}$ ($j = 1, 2, \dots, h_2$) simultaneously at time t . It follows from the above definitions that

$$\int f_{h_1 h_2} d\xi_{h_1 h_2} = n_{h_1 h_2}. \quad (1)$$

Since the quotient $\frac{n_{h_1+1, h_2} d\mathbf{x}_1^{(h_1+1)}}{n_{h_1 h_2}}$ represents the probability of finding a molecule of the first kind at $\mathbf{x}_1^{(h_1+1)}, d\mathbf{x}_1^{(h_1+1)}$ knowing that $d\mathbf{x}_{h_1 h_2}$ is occupied by h_1 molecules of the first kind and h_2 of the second kind, one has

$$\int n_{h_1+1, h_2} d\mathbf{x}_1^{(h_1+1)} = (N_1 - h_1) n_{h_1 h_2}. \quad (2)$$

$$\text{Similarly,} \quad \int f_{h_1+1, h_2} d\mathbf{x}_1^{(h_1+1)} d\boldsymbol{\xi}_1^{(h_1+1)} = (N_1 - h_1) f_{h_1 h_2}. \quad (3)$$

By repeating the above process it can be shown that

$$\left. \begin{aligned} \int n_{h_1 h_2} d\mathbf{x}_{h_1 h_2} &= \frac{N_1! N_2!}{(N_1 - h_1)! (N_2 - h_2)!}, \\ \iint f_{h_1 h_2} d\mathbf{x}_{h_1 h_2} d\boldsymbol{\xi}_{h_1 h_2} &= \frac{N_1! N_2!}{(N_1 - h_1)! (N_2 - h_2)!} \end{aligned} \right\} \quad (4)$$

When either, one of h_1 and h_2 is zero and the other is one, single-index notation will be used, e.g. n_{10} and f_{01} will be replaced by n_1 and f_2 respectively. In accordance with the usual notation, n_1 and n_2 are the number densities, and f_1 and f_2 the velocity distribution functions of the two kinds of molecules normalized with respect to their number densities. The partial mass densities are $\rho_1 = m_1 n_1$ and $\rho_2 = m_2 n_2$. The total number and mass density are $n = n_1 + n_2$ and $\rho = \rho_1 + \rho_2$ respectively.

3. FUNDAMENTAL CONSERVATION EQUATIONS

All fundamental conservation equations will be derived from Liouville's theorem in classical phase space of the system of molecules considered in § 2. Before deriving these equations a number of quantities pertaining to the set of molecules h_1, h_2 will be introduced; the Hamiltonian $H(h_1, h_2)$ for the set of molecules h_1, h_2 is given by

$$H(h_1 h_2) = K(h_1 h_2) + W(h_1 h_2) + E(h_1 h_2), \quad (5)$$

$$\left. \begin{aligned} \text{where} \quad K(h_1 h_2) &= \frac{1}{2} \left(\sum_{i=1}^{h_1} m_1 (\boldsymbol{\xi}_1^{(i)})^2 + \sum_{j=1}^{h_2} m_2 (\boldsymbol{\xi}_2^{(j)})^2 \right), \\ W(h_1 h_2) &= \frac{1}{2} \sum_{i,k=1}^{h_1} \phi_1^{(ik)} + \frac{1}{2} \sum_{j,l=1}^{h_2} \phi_2^{(jl)} + \sum_{i=1}^{h_1} \sum_{j=1}^{h_2} \psi^{(ij)}, \\ E(h_1 h_2) &= \sum_{i=1}^{h_1} E_1^{(i)} + \sum_{j=1}^{h_2} E_2^{(j)}. \end{aligned} \right\} \quad (6)$$

$\phi_1^{(ik)} = \phi_1(|\mathbf{x}_1^{(i)} - \mathbf{x}_1^{(k)}|)$ is the potential energy of two molecules of the first kind, $\phi_2^{(jl)} = \phi_2(|\mathbf{x}_2^{(j)} - \mathbf{x}_2^{(l)}|)$ that of two molecules of the second kind and $\psi^{(ij)} = \psi(|\mathbf{x}_1^{(i)} - \mathbf{x}_2^{(j)}|)$ that of two molecules of different kinds, it being assumed that the intermolecular forces are central and depend only on the distances between the mass centres of the molecules. $E_1^{(i)} = E_1(\mathbf{x}_1^{(i)})$ is the potential energy of a molecule of the first kind at $\mathbf{x}_1^{(i)}$ in the external force field assumed to be conservative and $E_2^{(j)} = E_2(\mathbf{x}_2^{(j)})$ that of a molecule of the second kind at $\mathbf{x}_2^{(j)}$.

The drift of $f_{N_1 N_2}$ in phase space is governed by the Liouville's theorem

$$\frac{\partial}{\partial t} f_{N_1 N_2} = [H(N_1 N_2), f_{N_1 N_2}], \quad (7)$$

where the square bracket $[]$ is the classical Poisson bracket defined by

$$[\alpha, \beta] = \frac{1}{m_1} \sum_{i=1}^{N_1} \left(\frac{\partial \alpha}{\partial \mathbf{x}_1^{(i)}} \frac{\partial \beta}{\partial \boldsymbol{\xi}_1^{(i)}} - \frac{\partial \alpha}{\partial \boldsymbol{\xi}_1^{(i)}} \frac{\partial \beta}{\partial \mathbf{x}_1^{(i)}} \right) + \frac{1}{m_2} \sum_{j=1}^{N_2} \left(\frac{\partial \alpha}{\partial \mathbf{x}_2^{(j)}} \frac{\partial \beta}{\partial \boldsymbol{\xi}_2^{(j)}} - \frac{\partial \alpha}{\partial \boldsymbol{\xi}_2^{(j)}} \frac{\partial \beta}{\partial \mathbf{x}_2^{(j)}} \right). \quad (8)$$

By integrating both sides of (7) over the position and velocity co-ordinates of all the molecules except the set h_1, h_2 , one obtains the equation of motion for $f_{h_1 h_2}$

$$\frac{\partial}{\partial t} f_{h_1 h_2} = [H(h_1 h_2), f_{h_1 h_2}] + \iint [V_{h_1 h_2}^{(h_1+1)}, f_{h_1+1, h_2}] d\mathbf{x}_1^{(h_1+1)} d\xi_1^{(h_1+1)} \\ + \iint [V_{h_1 h_2}^{(h_2+1)}, f_{h_1, h_2+1}] d\mathbf{x}_2^{(h_2+1)} d\xi_2^{(h_2+1)}, \quad (9)$$

where

$$V_{h_1 h_2}^{(h_1+1)} = \left\{ \sum_{i=1}^{h_1} \phi_1^{(i, h_1+1)} + \sum_{j=1}^{h_2} \psi^{(j, h_1+1)} \right\} \\ V_{h_1 h_2}^{(h_2+1)} = \left\{ \sum_{j=1}^{h_2} \phi_2^{(j, h_2+1)} + \sum_{i=1}^{h_1} \psi^{(i, h_2+1)} \right\} \quad (10)$$

In the following i and k will be reserved for indicating quantities pertaining to the first kind of molecules, j and l for the second kind of molecules, and no mention will be made of the kind of molecules concerned.

Let $\varphi_1^{(i)} = \varphi_1(t, \mathbf{x}_1^{(i)}, \xi_1^{(i)})$ be any molecular property of a molecule at $\mathbf{x}_1^{(i)}$ with velocity $\xi_1^{(i)}$. $\varphi_1^{(i)}$ may be a scalar, a vector or a tensor. Multiplying (9) by $\varphi_1^{(i)} d\xi_{h_1 h_2}$ and integrating over the velocities, we shall show that for $\varphi_1^{(i)} = 1$, $\xi_1^{(i)}$ and $\frac{1}{2}\xi^{(i)2}$, the equation of continuity, of conservation of momentum and energy of the set h_1, h_2 will follow.

For $\varphi_1^{(i)} = 1$, one obtains

$$\frac{\partial}{\partial t} n_{h_1 h_2} = \int [K(h_1 h_2), f_{h_1 h_2}] d\xi_{h_1 h_2}, \quad (11)$$

since $W(h_1 h_2)$, $E(h_1 h_2)$, $V_{h_1 h_2}^{(h_1+1)}$ and $V_{h_1 h_2}^{(h_2+1)}$ are functions of position co-ordinates only, and the integrals involving these quantities can be transformed into vanishing surface integrals. If one defines $\mathbf{u}_{h_1 h_2}^{(i)}$ by

$$\mathbf{u}_{h_1 h_2}^{(i)} = \frac{1}{n_{h_1 h_2}} \int f_{h_1 h_2} \xi_1^{(i)} d\xi_{h_1 h_2} \quad (12)$$

with a similar definition for $\mathbf{u}_{h_1 h_2}^{(j)}$, equation (11) becomes

$$\frac{\partial}{\partial t} n_{h_1 h_2} + \sum_{i=1}^{h_1} \frac{\partial}{\partial \mathbf{x}_1^{(i)}} \cdot (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(i)}) + \sum_{j=1}^{h_2} \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \cdot (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(j)}) = 0, \quad (13)$$

which is the equation of continuity for the set h_1, h_2 , $\mathbf{u}_{h_1 h_2}^{(i)}$ is the average velocity of a molecule at $\mathbf{x}_1^{(i)}$ in the set h_1, h_2 whose positions are specified. By introducing the convective time derivative d/dt defined by

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \sum_{i=1}^{h_1} \mathbf{u}_{h_1 h_2}^{(i)} \cdot \frac{\partial}{\partial \mathbf{x}_1^{(i)}} + \sum_{j=1}^{h_2} \mathbf{u}_{h_1 h_2}^{(j)} \cdot \frac{\partial}{\partial \mathbf{x}_2^{(j)}}, \quad (14)$$

equation (13) becomes

$$\frac{d}{dt} n_{h_1 h_2} + n_{h_1 h_2} \left(\sum_{i=1}^{h_1} \frac{\partial}{\partial \mathbf{x}_1^{(i)}} \cdot \mathbf{u}_{h_1 h_2}^{(i)} + \sum_{j=1}^{h_2} \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \cdot \mathbf{u}_{h_1 h_2}^{(j)} \right) = 0. \quad (15)$$

For $\varphi_1^{(i)} = \xi_1^{(i)}$, one obtains after performing partial integrations on integrals involving $W(h_1 h_2)$, $E(h_1 h_2)$, $V_{h_1 h_2}^{(h_1+1)}$ and $V_{h_1 h_2}^{(h_2+1)}$,

$$\frac{\partial}{\partial t} (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(i)}) = \int [K(h_1 h_2), f_{h_1 h_2}] \xi_1^{(i)} d\xi_{h_1 h_2} + n_{h_1 h_2} \gamma_{h_1 h_2}^{(i)}, \quad (16)$$

where

$$\mathbf{r}_{h_1 h_2}^{(i)} = -\frac{1}{m_1} \frac{\partial}{\partial \mathbf{x}_1^{(i)}} (W(h_1 h_2) + E(h_1 h_2)) - \frac{1}{m_1} \left\{ \int \frac{n_{h_1+1, h_2}}{n_{h_1 h_2}} \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} d\mathbf{x}_1^{(h_1+1)} + \int \frac{n_{h_1, h_2+1}}{n_{h_1 h_2}} \frac{\partial \psi^{(i, h_2+1)}}{\partial \mathbf{x}_1^{(i)}} d\mathbf{x}_2^{(h_2+1)} \right\}. \quad (17)$$

It follows from (13) and (14) that

$$\frac{\partial}{\partial t} (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(i)}) = n_{h_1 h_2} \frac{d}{dt} \mathbf{u}_{h_1 h_2}^{(i)} - \left\{ \sum_{k=1}^{h_1} \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(k)} \mathbf{u}_{h_1 h_2}^{(i)}) + \sum_{l=1}^{h_2} \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(l)} \mathbf{u}_{h_1 h_2}^{(i)}) \right\}. \quad (18)$$

Combining (18) with (16), one obtains

$$\frac{d}{dt} (m_1 \mathbf{u}_{h_1 h_2}^{(i)}) + \frac{1}{n_{h_1 h_2}} \left(\sum_k \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{K}_{h_1 h_2}^{(ki)} + \sum_l \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{K}_{h_1 h_2}^{(li)} \right) = m_1 \mathbf{r}_{h_1 h_2}^{(i)}, \quad (19)$$

where
$$\mathbf{K}_{h_1 h_2}^{(ki)} = m_1 \int f_{h_1 h_2} \mathbf{v}_{h_1 h_2}^{(k)} \mathbf{v}_{h_1 h_2}^{(i)} d\mathbf{x}_{h_1 h_2} \quad (20)$$

with a similar expression for $\mathbf{K}_{h_1 h_2}^{(li)}$.

An equation similar to (19) can be written for the rate of change of $\mathbf{u}_{h_1 h_2}^{(j)}$,

$$\frac{d}{dt} (m_2 \mathbf{u}_{h_1 h_2}^{(j)}) + \frac{1}{n_{h_1 h_2}} \left(\sum_k \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{K}_{h_1 h_2}^{(kj)} + \sum_l \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{K}_{h_1 h_2}^{(lj)} \right) = m_2 \mathbf{r}_{h_1 h_2}^{(j)}, \quad (21)$$

where
$$\mathbf{K}_{h_1 h_2}^{(kj)} = m_2 \int f_{h_1 h_2} \mathbf{v}_{h_1 h_2}^{(k)} \mathbf{v}_{h_1 h_2}^{(j)} d\mathbf{x}_{h_1 h_2}, \quad (22)$$

and $\mathbf{K}_{h_1 h_2}^{(lj)}$ is similarly defined.

One can bring (19) and (20) into still simpler form by introducing $\mathbf{L}_{h_1 h_2}^{(ki)}$ and $\mathbf{L}_{h_1 h_2}^{(li)}$ defined by the equations

$$\left. \begin{aligned} \sum_k \left(\frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{L}_{h_1 h_2}^{(ki)} \right) &= n_{h_1 h_2} \left(\sum_k \frac{\partial \phi_1^{(i, k)}}{\partial \mathbf{x}_1^{(i)}} + \int \frac{n_{h_1+1, h_2}}{n_{h_1 h_2}} \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} d\mathbf{x}_1^{(h_1+1)} \right), \\ \sum_l \left(\frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{L}_{h_1 h_2}^{(li)} \right) &= n_{h_1 h_2} \left(\sum_l \frac{\partial \psi^{(i, l)}}{\partial \mathbf{x}_2^{(i)}} + \int \frac{n_{h_1, h_2+1}}{n_{h_1 h_2}} \frac{\partial \psi^{(i, h_2+1)}}{\partial \mathbf{x}_2^{(i)}} d\mathbf{x}_2^{(h_2+1)} \right). \end{aligned} \right\} \quad (23)$$

Equations (19) and (20) then become

$$\left. \begin{aligned} \frac{d}{dt} (m_1 \mathbf{u}_{h_1 h_2}^{(i)}) + \frac{1}{n_{h_1 h_2}} \left\{ \sum_k \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{P}_{h_1 h_2}^{(ki)} + \sum_l \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{P}_{h_1 h_2}^{(li)} \right\} &= \mathbf{F}_1^{(i)}, \\ \frac{d}{dt} (m_2 \mathbf{u}_{h_1 h_2}^{(j)}) + \frac{1}{n_{h_1 h_2}} \left\{ \sum_k \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{P}_{h_1 h_2}^{(kj)} + \sum_l \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{P}_{h_1 h_2}^{(lj)} \right\} &= \mathbf{F}_2^{(j)}, \end{aligned} \right\} \quad (24)$$

where
$$\left. \begin{aligned} \mathbf{P}_{h_1 h_2}^{(ki)} &= \mathbf{K}_{h_1 h_2}^{(ki)} + \mathbf{L}_{h_1 h_2}^{(ki)}, \quad \text{etc.}, \\ \mathbf{F}_1^{(i)} &= -\frac{\partial}{\partial \mathbf{x}_1^{(i)}} E_1^{(i)}, \quad \text{etc.} \end{aligned} \right\} \quad (25)$$

The interpretation of these equations is clear. Equation (24) gives the rate of change of the average velocity of a molecule in the set h_1, h_2 whose positions are specified; apart from the external force acting on this molecule, there are the kinetic

pressure force $\frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{K}_{h_1 h_2}^{(ki)}$, etc., defined in a similar manner as in gas theory, and the potential pressure force $\frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{L}_{h_1 h_2}^{(ki)}$, etc., which consists of forces due to the rest of the set h_1, h_2 whose positions are specified and the set $(N_1 - h_1), (N_2 - h_2)$ whose positions are not specified.

One can interpret the pressure tensor even more naturally by considering the motion of a representative point in the $3(h_1 + h_2)$ dimensional phase space. Let $\mathbf{x}_{h_1 h_2}$ be a vector in the $3(h_1 + h_2)$ -space whose components are $\mathbf{x}_1^{(i)}$ ($i = 1, 2, \dots, h_1$) and $\mathbf{x}_2^{(j)}$ ($j = 1, 2, \dots, h_2$), $\mathbf{G}_{h_1 h_2}$ the momentum vector whose components are $m_1 \mathbf{u}_{h_1 h_2}^{(i)}$ and $m_2 \mathbf{u}_{h_1 h_2}^{(j)}$, $\mathbf{F}_{h_1 h_2}$ the external force vector whose components are $\mathbf{F}_1^{(i)}$ and $\mathbf{F}_2^{(j)}$, $\mathbf{K}_{h_1 h_2}$ the kinetic pressure tensor whose components are $\mathbf{K}_{h_1 h_2}^{(ki)}$, $\mathbf{K}_{h_1 h_2}^{(li)}$, $\mathbf{K}_{h_1 h_2}^{(kj)}$ and $\mathbf{K}_{h_1 h_2}^{(lj)}$ ($i, k = 1, 2, \dots, h_1; j, l = 1, 2, \dots, h_2$) and similarly for $\mathbf{L}_{h_1 h_2}$. The total pressure tensor is

$$\mathbf{P}_{h_1 h_2} = \mathbf{K}_{h_1 h_2} + \mathbf{L}_{h_1 h_2}. \quad (26)$$

One has then from (24)

$$\frac{d}{dt} \mathbf{G}_{h_1 h_2} + \frac{1}{n_{h_1 h_2}} \left\{ \frac{\partial}{\partial \mathbf{x}_{h_1 h_2}} \cdot \mathbf{P}_{h_1 h_2} \right\} = \mathbf{F}_{h_1 h_2}, \quad (27)$$

which is an exact analogue of ordinary hydrodynamical equation except for the difference in the number of dimensions.

For $\varphi_1^{(i)} = \frac{1}{2} \xi_1^{(i)2}$, one obtains, using (17),

$$\begin{aligned} & \frac{\partial}{\partial t} \left\{ \frac{1}{2} n_{h_1 h_2} (\mathbf{u}_{h_1 h_2}^{(i)})^2 + \frac{3k}{2m_1} n_{h_1 h_2} T_{h_1 h_2}^{(i)} \right\} \\ &= \int [K(h_1 h_2), f_{h_1 h_2}] \frac{1}{2} \xi_1^{(i)2} d\xi_{h_1 h_2} + n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(i)} \cdot \boldsymbol{\eta}_{h_1 h_2}^{(i)} - \frac{1}{m_1} \\ & \times \left\{ \int \frac{\partial p_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} \cdot (\mathbf{u}_{h_1+1, h_2}^{(i)} - \mathbf{u}_{h_1 h_2}^{(i)}) n_{h_1+1, h_2} d\mathbf{x}_1^{(h_1+1)} \right. \\ & \left. + \int \frac{\partial \psi^{(i, h_2+1)}}{\partial \mathbf{x}_1^{(i)}} \cdot (\mathbf{u}_{h_1, h_2+1}^{(i)} - \mathbf{u}_{h_1 h_2}^{(i)}) n_{h_1, h_2+1} d\mathbf{x}_2^{(h_2+1)} \right\}, \quad (28) \end{aligned}$$

where

$$\frac{3}{2} k T_{h_1 h_2}^{(i)} = \frac{1}{n_{h_1 h_2}} \int f_{h_1 h_2} \frac{1}{2} m_1 v_{h_1 h_2}^{(i)2} d\xi_{h_1 h_2} \quad (29)$$

defines the generalized temperature of a molecule at $\mathbf{x}_1^{(i)}$ in the set h_1, h_2 . With the help of (14) and (19), equation (28) can be transformed to the form

$$\begin{aligned} & \frac{3}{2} k n_{h_1 h_2} \frac{d}{dt} T_{h_1 h_2}^{(i)} + \sum_k \left\{ \mathbf{K}_{h_1 h_2}^{(ki)} \cdot \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \mathbf{u}_{h_1 h_2}^{(i)} + \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{m}_{h_1 h_2}^{(ki)} \right\} + \sum_l \left\{ \mathbf{K}_{h_1 h_2}^{(li)} \cdot \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \mathbf{u}_{h_1 h_2}^{(i)} + \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{m}_{h_1 h_2}^{(li)} \right\} \\ &= - \left\{ \int n_{h_1+1, h_2} \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} \cdot (\mathbf{u}_{h_1+1, h_2}^{(i)} - \mathbf{u}_{h_1 h_2}^{(i)}) d\mathbf{x}_1^{(h_1+1)} \right. \\ & \left. + \int n_{h_1, h_2+1} \frac{\partial \psi^{(i, h_2+1)}}{\partial \mathbf{x}_1^{(i)}} \cdot (\mathbf{u}_{h_1, h_2+1}^{(i)} - \mathbf{u}_{h_1 h_2}^{(i)}) d\mathbf{x}_2^{(h_2+1)} \right\}, \quad (30) \end{aligned}$$

where

$$\mathbf{m}_{h_1 h_2}^{(ki)} = m_1 \int f_{h_1 h_2} \mathbf{v}_{h_1 h_2}^{(k)} \frac{1}{2} (v_{h_1 h_2}^{(i)})^2 d\xi_{h_1 h_2}, \quad (31)$$

and $\mathbf{m}_{h_1 h_2}^{(li)}$ is similarly defined.

The external force $F_1^{(i)}$ which affects only the mass-motion velocity and the internal forces due to the rest of the set h_1, h_2 do not enter into (30) as one would expect. By strict analogy with the gas theory one sees that the vectors $\mathbf{m}_{h_1 h_2}^{(ki)}$, $\mathbf{m}_{h_1 h_2}^{(li)}$ are together the generalized kinetic thermal energy flux connected with a molecule at $\mathbf{x}_1^{(i)}$ in the set h_1, h_2 .

An equation similar to (30) can be written for the rate of change of $T_{h_1 h_2}^{(j)}$. To find the corresponding equation for the rate of change of the internal potential energy, ones writes $I_{h_1 h_2}^{(i)}$ for the internal potential energy of a molecule at $\mathbf{x}_1^{(i)}$ in the set h_1, h_2 :

$$I_{h_1 h_2}^{(i)} = \frac{1}{2} \sum_k \phi_1^{(ik)} + \frac{1}{2} \sum_l \psi^{(il)} + \frac{1}{2} \int \frac{n_{h_1+1, h_2}}{n_{h_1 h_2}} \phi_1^{(i, h_1+1)} d\mathbf{x}_1^{(h_1+1)} + \frac{1}{2} \int \frac{n_{h_1, h_2+1}}{n_{h_1 h_2}} \psi^{(i, h_2+1)} d\mathbf{x}_2^{(h_2+1)}, \quad (32)$$

where the factor $\frac{1}{2}$ signifies that the mutual potential energy is shared by each pair. Using (13) and (14), one has for the rate of change of $I_{h_1 h_2}^{(i)}$

$$\begin{aligned} n_{h_1 h_2} \frac{d}{dt} I_{h_1 h_2}^{(i)} + \sum_k \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{n}_{h_1 h_2}^{(ki)} + \sum_l \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{n}_{h_1 h_2}^{(li)} \\ = \int n_{h_1+1, h_2} \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} \cdot (\mathbf{u}_{h_1+1, h_2}^{(i)} - \mathbf{u}_{h_1 h_2}^{(i)}) d\mathbf{x}_1^{(h_1+1)} \\ + \int n_{h_1, h_2+1} \frac{\partial \psi^{(i, h_2+1)}}{\partial \mathbf{x}_1^{(i)}} \cdot (\mathbf{u}_{h_1, h_2+1}^{(i)} - \mathbf{u}_{h_1 h_2}^{(i)}) d\mathbf{x}_2^{(h_2+1)}, \quad (33) \end{aligned}$$

where

$$\begin{aligned} \sum_k \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{n}_{h_1 h_2}^{(ki)} = \frac{1}{2} \sum_k \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \left\{ \int n_{h_1+1, h_2} \phi_1^{(i, h_1+1)} (\mathbf{u}_{h_1+1, h_2}^{(k)} - \mathbf{u}_{h_1 h_2}^{(k)}) d\mathbf{x}_1^{(h_1+1)} \right. \\ \left. + \int n_{h_1, h_2+1} \psi^{(i, h_2+1)} (\mathbf{u}_{h_1, h_2+1}^{(k)} - \mathbf{u}_{h_1 h_2}^{(k)}) d\mathbf{x}_2^{(h_2+1)} \right\} + \frac{1}{2} n_{h_1 h_2} \sum_k (\mathbf{u}_{h_1 h_2}^{(k)} - \mathbf{u}_{h_1 h_2}^{(i)}) \cdot \frac{\partial \phi_1^{(i, k)}}{\partial \mathbf{x}_1^{(i)}} \\ + \frac{1}{2} \int n_{h_1+1, h_2} (\mathbf{u}_{h_1+1, h_2}^{(h_1+1)} + \mathbf{u}_{h_1+1, h_2}^{(i)} - 2\mathbf{u}_{h_1 h_2}^{(i)}) \cdot \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} d\mathbf{x}_1^{(h_1+1)}, \quad (34) \end{aligned}$$

with a similar expression for $\sum_l \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{n}_{h_1 h_2}^{(li)}$. By combining (30) and (33), one has

$$n_{h_1 h_2} \frac{d}{dt} E_{h_1 h_2}^{(i)} + \sum_k \left\{ K_{h_1 h_2}^{(ki)} : \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \mathbf{u}_{h_1 h_2}^{(i)} + \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \cdot \mathbf{q}_{h_1 h_2}^{(ki)} \right\} + \sum_l \left\{ K_{h_1 h_2}^{(li)} : \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \mathbf{u}_{h_1 h_2}^{(i)} + \frac{\partial}{\partial \mathbf{x}_2^{(l)}} \cdot \mathbf{q}_{h_1 h_2}^{(li)} \right\} = 0, \quad (35)$$

where

$$\left. \begin{aligned} E_{h_1 h_2}^{(i)} &= \frac{3}{2} k T_{h_1 h_2}^{(i)} + I_{h_1 h_2}^{(i)}, \\ \mathbf{q}_{h_1 h_2}^{(ki)} &= \mathbf{m}_{h_1 h_2}^{(ki)} + \mathbf{n}_{h_1 h_2}^{(ki)}, \quad \text{etc.} \end{aligned} \right\} \quad (36)$$

The interpretation of (35) can be given in either of the two ways in which equation (24) has been interpreted. Each of $E_{h_1 h_2}^{(i)}$ and $\mathbf{q}_{h_1 h_2}^{(ki)}$ consists of two parts, the kinetic and the potential part, while the terms $\sum_k \left(K_{h_1 h_2}^{(ki)} : \frac{\partial}{\partial \mathbf{x}_1^{(k)}} \mathbf{u}_{h_1 h_2}^{(i)} \right)$, etc., represent the dissipation of mechanical energy into heat due to internal friction.

4. REDUCTION OF THE MOLECULAR CONSERVATION EQUATIONS TO ORDINARY HYDRODYNAMICAL EQUATIONS

In an assembly consisting of only one kind of molecules, the generalized definitions of the kinetic pressure, the temperature and the thermal flux reduce for $h = 1$ in each case to their respective ordinary definitions. This is not so in a mixture where the generalized definitions of the above quantities reduce for $h_1 = 1, h_2 = 0$ to quantities referring to the first kind of molecules only, which cannot be directly observed. What one directly observes in a mixture, instead of the above-mentioned quantities, are the average properties contributed by a small group of molecules of various types in a small but finite time interval. Nevertheless, it is possible to find simple connexions between the two sets of quantities so that either of them can be expressed in terms of the other.

There are two different average velocities in a mixture: the directly observable mass velocity \mathbf{u} relative to which the peculiar velocities (or thermal velocities) of the various types of molecules, and therefore the kinetic pressure, temperature and thermal flux in a mixture are defined, and the number velocity \mathbf{u}_n used in connexion with the transport of molecular properties

$$\mathbf{u} = \frac{\rho_1 \mathbf{u}_1 + \rho_2 \mathbf{u}_2}{\rho}, \quad \mathbf{u}_n = \frac{n_1 \mathbf{u}_1 + n_2 \mathbf{u}_2}{n}, \quad (37)$$

$$\text{where} \quad \mathbf{u}_1 = \int \frac{f_1}{n_1} \xi_1^{(1)} d\xi_1^{(1)}, \quad \mathbf{u}_2 = \int \frac{f_2}{n_2} \xi_2^{(1)} d\xi_2^{(1)}, \quad (38)$$

f_1, n_1 and $\xi_1^{(1)}$ stand for $f_{10}(\mathbf{x}_1^{(1)}, \xi_1^{(1)}, t)$, $n_{10}(\mathbf{x}_1^{(1)}, t)$ and $\xi_{10}^{(1)}$ respectively.

Let K, T and \mathbf{m} be the kinetic pressure, temperature and thermal flux in a mixture. These are defined as

$$\left. \begin{aligned} K &= K_1 + K_2 \\ K_1 &= m_1 \int f_1 \mathbf{v}_1 \mathbf{v}_1 d\mathbf{v}_1, \quad K_2 = m_2 \int f_2 \mathbf{v}_2 \mathbf{v}_2 d\mathbf{v}_2, \\ nT &= n_1 T_1 + n_2 T_2, \\ n_1 T_1 &= \frac{m_1}{3k} \int f_1 v_1^2 d\mathbf{v}_1, \quad n_2 T_2 = \frac{m_2}{3k} \int f_2 v_2^2 d\mathbf{v}_2, \\ \mathbf{m} &= \mathbf{m}_1 + \mathbf{m}_2, \\ \mathbf{m}_1 &= \frac{m_1}{2} \int f_1 \mathbf{v}_1 v_1^2 d\mathbf{v}_1, \quad \mathbf{m}_2 = \frac{m_2}{2} \int f_2 \mathbf{v}_2 v_2^2 d\mathbf{v}_2, \end{aligned} \right\} \quad (39)$$

$$\text{where} \quad \mathbf{v}_1 = \xi_1^{(1)} - \mathbf{u}, \quad \mathbf{v}_2 = \xi_2^{(1)} - \mathbf{u}. \quad (40)$$

K_1 and \mathbf{m}_1 are the partial kinetic pressure and the thermal flux of the molecules of the first kind, and T_1 the partial temperature such that $\frac{1}{2} k T_1$ is the average heat energy for each degree of freedom of a molecule of the first kind. K_2, \mathbf{m}_2 and T_2 have similar meanings for the second kind of molecules.

On the other hand, one obtains from (20), (29) and (31) for $h_1 = 1$, $h_2 = 0$

$$\left. \begin{aligned} K_{10} &= m_1 \int f_1 v_{10} v_{10} dv_{10}, \\ \frac{3}{2} k T_{10} &= \frac{m_1}{2} \int \frac{f_1}{n_1} v_{10}^2 dv_{10}, \\ \mathbf{m}_{10} &= \frac{m_1}{2} \int f_1 v_{10} v_{10}^2 dv_{10}, \end{aligned} \right\} \quad (41)$$

with similar expressions for K_{01} , T_{01} and \mathbf{m}_{01} , where $\mathbf{v}_{10} = \xi_1^{(1)} - \mathbf{u}_1$.

From the above definitions, it can be shown that

$$\left. \begin{aligned} K &= K_{10} + K_{01} + \rho_1 \mathbf{u}'_1 \mathbf{u}'_1 + \rho_2 \mathbf{u}'_2 \mathbf{u}'_2, \\ \frac{3}{2} n k T &= \frac{3}{2} k (n_1 T_{10} + n_2 T_{01}) + \frac{1}{2} \rho_1 (\mathbf{u}'_1)^2 + \frac{1}{2} \rho_2 (\mathbf{u}'_2)^2, \\ \mathbf{m} &= \mathbf{m}_{10} + \mathbf{m}_{01} + K_{10} \cdot \mathbf{u}'_1 + K_{01} \cdot \mathbf{u}'_2 + \frac{3}{2} k T (n_1 \mathbf{u}'_1 + n_2 \mathbf{u}'_2), \end{aligned} \right\} \quad (42)$$

where

$$\mathbf{u}'_1 = \mathbf{u}_1 - \mathbf{u}, \quad \mathbf{u}'_2 = \mathbf{u}_2 - \mathbf{u}. \quad (43)$$

With these relations between K , T , \mathbf{m} and K_{10} , T_{10} , \mathbf{m}_{10} , etc., it can be shown that the conservation equations (13), (24) and (30) reduce for $h_1 = 1$, $h_2 = 0$ or $h_1 = 0$, $h_2 = 1$ to equations similar in form to the corresponding equations in a gas mixture (Chapman & Cowling 1939), these are

$$\left. \begin{aligned} \frac{\partial}{\partial t} n_1 &= - \frac{\partial}{\partial \mathbf{x}} \cdot (n_1 \mathbf{u}_1), \\ \frac{\partial}{\partial t} \mathbf{u} &= - \mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} \mathbf{u} - \frac{1}{\rho} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{K} + \frac{1}{\rho} (\rho_1 \boldsymbol{\eta}_1 + \rho_2 \boldsymbol{\eta}_2), \\ \frac{\partial}{\partial t} T &= - \mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} T + \frac{T}{n} \frac{\partial}{\partial \mathbf{x}} \cdot (n_1 \mathbf{u}'_1 + n_2 \mathbf{u}'_2) - \frac{2}{3 k n} \left(\mathbf{K} : \frac{\partial}{\partial \mathbf{x}} \mathbf{u} + \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{m} \right) \\ &\quad + \frac{2}{3 k n} \{ (\rho_1 \boldsymbol{\eta}_1 \cdot \mathbf{u}'_1 + \rho_2 \boldsymbol{\eta}_2 \cdot \mathbf{u}'_2) - (\delta_1 + \delta_2) \}, \end{aligned} \right\} \quad (44)$$

$$\text{where} \quad \rho_1 \boldsymbol{\eta}_1 = n_1 \mathbf{F}_1 - \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}_1, \quad \mathbf{F}_1 = \mathbf{F}^{(1)},$$

$$\left. \begin{aligned} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}_1 &= \int \frac{\partial \phi_1}{\partial \mathbf{x}_1^{(1)}} n_{20} d\mathbf{x}_1^{(2)} + \int \frac{\partial \psi}{\partial \mathbf{x}_1^{(1)}} n_{11} d\mathbf{x}_2^{(1)}, \\ \delta_1 &= \int n_{20} \frac{\partial \phi_1}{\partial \mathbf{x}_1^{(1)}} \cdot (\mathbf{u}_{20}^{(1)} - \mathbf{u}_1) d\mathbf{x}_1^{(2)} + \int n_{11} \frac{\partial \psi}{\partial \mathbf{x}_1^{(1)}} \cdot (\mathbf{u}_{11}^{(1)} - \mathbf{u}_1) d\mathbf{x}_2^{(1)}. \end{aligned} \right\} \quad (45)$$

n_{11} stands for $n_{11}(\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, t)$ and $\mathbf{u}_{11}^{(1)}$ is the average velocity of a molecule of the first kind at $\mathbf{x}_1^{(1)}$ in the presence of a molecule of the second kind at $\mathbf{x}_2^{(1)}$.

To compare the equations in (44) with the corresponding equations in the theory of gas mixtures, one notices that the part played by the divergence of the partial pressure corresponds to the rate of change of momentum of a molecule in a gas mixture due to collisions with other molecules. The condition that the total momentum of the molecules in a small volume element be unaltered by collisions among

them—a consequence of binary encounter and the hypothesis of molecular chaos—is no longer satisfied in the general case, but the rate of change of momentum is represented by the quantity

$$\frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}_1 + \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}_2 = \left\{ \int \frac{\partial \phi_1}{\partial \mathbf{x}_1^{(1)}} n_{20} d\mathbf{x}_1^{(2)} + \int \frac{\partial \psi}{\partial \mathbf{x}_1^{(1)}} n_{11} d\mathbf{x}_2^{(1)} \right\} + \left\{ \int \frac{\partial \phi_2}{\partial \mathbf{x}_2^{(1)}} n_{02} d\mathbf{x}_2^{(1)} + \int \frac{\partial \psi}{\partial \mathbf{x}_2^{(1)}} n_{11} d\mathbf{x}_2^{(1)} \right\},$$

which does not vanish owing to the interaction of a large number of molecules other than those under immediate consideration.

Similarly, the part played in the potential transport of thermal energy corresponds to the rate of change of the thermal energy of a given molecule due to collisions with other molecules. The condition that the total energy of the molecules in a small volume element be unaltered by collisions among them—again the consequence of binary encounters and the hypothesis of molecular chaos—is no longer satisfied in the general case, but the rate of change of thermal energy is represented by the quantity

$$-\left\{ \mathbf{u}'_1 \cdot \left(\frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}_1 \right) + \mathbf{u}'_2 \cdot \left(\frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}_2 \right) - (\delta'_1 + \delta'_2) \right\},$$

which does vanish owing to the interaction of a large number of molecules other than those at the point considered.

5. EXPANSION OF THE EQUATION OF MOTION

The equation of motion (9) as it stands possesses solutions far too general for our purpose. Following Born & Green, we shall in the following expand equation (9) in powers of the gradients of certain physical parameters describing the macroscopic state of a liquid mixture. Improvement has been made on the choice of parameters; the potential energies of the two types of molecules in the external force field (E_1, E_2), instead of the external forces ($\mathbf{F}_1, \mathbf{F}_2$), are taken to be on equal footing with the partial densities, the temperature and the mass-motion velocity, in order that the force diffusion should follow naturally after expansion.

The normal solution which we require deviates but slightly from the equilibrium solution and therefore can be separated into parts

$$f_{h_1 h_2} = f_{h_1 h_2}^0 + f'_{h_1 h_2} + f''_{h_1 h_2} + \dots \quad (46)$$

such that $f_{h_1 h_2}^0$, the zeroth approximation to $f_{h_1 h_2}$, is that equilibrium solution for a homogeneous fluid with properly chosen constant parameters which fits best the real $f_{h_1 h_2}$ in the region considered, $f'_{h_1 h_2}$, the difference of the first approximation to $f_{h_1 h_2}$ from $f_{h_1 h_2}^0$, is the correction for a non-uniform fluid with properly chosen constant gradients of the parameters to fit the real $f_{h_1 h_2}$, etc. A natural and convenient basis for choosing the values of these parameters and their gradients is to identify them with the values at the centre of gravity of the set h_1, h_2 . Hence we postulate

$$\begin{aligned} f_{h_1 h_2} &= F_{h_1 h_2} \left(\mathbf{r}_1^{(ik)}, \mathbf{r}_2^{(j)}, \mathbf{s}^{(ij)}, \mathbf{v}_1^{(i)}, \mathbf{v}_2^{(j)}, \lambda_i, \frac{\partial \lambda_i}{\partial \mathbf{x}}, \frac{\partial^2 \lambda}{\partial \mathbf{x} \partial \mathbf{x}}, \frac{\partial \lambda_i}{\partial \mathbf{x}}, \frac{\partial \lambda_{i_0}}{\partial \mathbf{x}}, \dots \right) \\ &= F_{h_1 h_2}^0 + F'_{h_1 h_2} + F''_{h_1 h_2} + \dots, \end{aligned} \quad (47)$$

where the capital $F_{h_1 h_2}$ is to denote the new functional dependence. $F_{h_1 h_2}^0$ does not involve any gradients, $F'_{h_1 h_2}$ involves gradients of the first order and first degree,

$F''_{h_1 h_2}$ involves gradients of the second order and first degree and also gradients of the first order to the second degree, etc.; λ_l , $\partial \lambda_l / \partial \mathbf{x}$, $\partial^2 \lambda_l / \partial \mathbf{x} \partial \mathbf{x}$, etc. ($l = 1, 2, \dots, 8$) represent the values of n_1 , n_2 , \mathbf{u} , T , E_1 , E_2 and their space gradients of all orders at the mass centre $\mathbf{x} = \left(\sum_i m_1 \mathbf{x}_1^{(i)} + \sum_j m_2 \mathbf{x}_2^{(j)} \right) / (h_1 m_1 + h_2 m_2)$ of the set h_1, h_2 . The quantities $\mathbf{r}_1^{(ik)}$, $\mathbf{r}_2^{(jl)}$ and $\mathbf{s}^{(ij)}$ are the distances between molecules, defined by $\mathbf{r}_1^{(ik)} = \mathbf{x}_1^{(k)} - \mathbf{x}_1^{(i)}$, $\mathbf{r}_2^{(jl)} = \mathbf{x}_2^{(l)} - \mathbf{x}_2^{(j)}$ and $\mathbf{s}^{(ij)} = \mathbf{x}_2^{(j)} - \mathbf{x}_1^{(i)}$. That the dependence on $\mathbf{x}_1^{(i)}$, $\mathbf{x}_2^{(j)}$ of $F_{h_1 h_2}$, appears in the combinations $\mathbf{r}_1^{(ik)}$, $\mathbf{r}_2^{(jl)}$, $\mathbf{s}^{(ij)}$ is suggested by the fact that intermolecular forces can be represented as functions of the distance between the mass centres of the pair of molecules (at least for mono-atomic substances). As a consequence, one has

$$\left(\sum_i \frac{\partial}{\partial \mathbf{x}_1^{(i)}} + \sum_j \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \right) F_{h_1 h_2} = \sum_{i,k} \frac{\partial F_{h_1 h_2}}{\partial \mathbf{r}_1^{(ik)}} + \sum_{j,l} \frac{\partial F_{h_1 h_2}}{\partial \mathbf{r}_2^{(jl)}} + \sum_i \sum_j \left(\frac{\partial F_{h_1 h_2}}{\partial \mathbf{s}^{(ij)}} + \frac{\partial F_{h_1 h_2}}{\partial \mathbf{s}^{(ji)}} \right) = 0. \quad (48)$$

The dependence of $F_{h_1 h_2}$ on \mathbf{x} and t is implicit in λ_l , $\partial \lambda_l / \partial \mathbf{x}$, etc. As λ_l involves \mathbf{u} , the dependence of $F_{h_1 h_2}$ on $\xi_1^{(i)}$ and $\xi_2^{(j)}$ can be made in terms of $\mathbf{v}_1^{(i)} = \xi_1^{(i)} - \mathbf{u}$ and $\mathbf{v}_2^{(j)} = \xi_2^{(j)} - \mathbf{u}$.

The above considerations hold as well for all quantities derivable from $F_{h_1 h_2}$, such as $N_{h_1 h_2}$, $\mathbf{U}_{h_1 h_2}^{(i)}$, $T_{h_1 h_2}^{(i)}$, $K_{h_1 h_2}^{(ik)}$, etc. (A capital letter is used in each case to indicate its new functional dependence.) Also their space gradients and time derivatives may be expanded in a similar way provided

$$\frac{\partial \lambda_l}{\partial t} = \left(\frac{\partial \lambda_l}{\partial t} \right)^0 + \left(\frac{\partial \lambda_l}{\partial t} \right)' + \left(\frac{\partial \lambda_l}{\partial t} \right)'' + \dots \quad (49)$$

is known. It will be shown in the following that $(\partial \lambda_l / \partial t)^0 = 0$ for all λ_l and $(\partial \lambda_l / \partial t)'$ can be calculated from $f_{h_1 h_2}^0$, $(\partial \lambda_l / \partial t)''$ from $f_{h_1 h_2}'$, etc.

When $f_{h_1 h_2}^0$, $f_{h_1 h_2}'$ are found, any of the quantities derivable from $f_{h_1 h_2}$ can be calculated to the corresponding degree of approximation, for example:

$$\left. \begin{aligned} n_{h_1 h_2}^0 &= \int f_{h_1 h_2}^0 d\xi_{h_1 h_2}, & n_{h_1 h_2}' &= \int f_{h_1 h_2}' d\xi_{h_1 h_2}, \\ \mathbf{u}_{h_1 h_2}^{(i)0} &= \int \frac{f_{h_1 h_2}^0}{n_{h_1 h_2}^0} \xi_1^{(i)} d\xi_{h_1 h_2}, & \mathbf{u}_{h_1 h_2}^{(i)'} &= \int \frac{f_{h_1 h_2}'}{n_{h_1 h_2}^0} \mathbf{v}_{h_1 h_2}^{(i)0} d\xi_{h_1 h_2}, \\ \mathbf{v}_{h_1 h_2}^{(i)0} &= \xi_1^{(i)} - \mathbf{u}_{h_1 h_2}^{(i)0}, & \mathbf{v}_{h_1 h_2}^{(i)'} &= -\mathbf{u}_{h_1 h_2}^{(i)'} \\ T_{h_1 h_2}^{(i)0} &= \frac{m_1}{3k} \int \frac{f_{h_1 h_2}^0}{n_{h_1 h_2}^0} (\mathbf{v}_{h_1 h_2}^{(i)0})^2 d\xi_{h_1 h_2}, & T_{h_1 h_2}^{(i)'} &= \frac{m_1}{3k} \int \frac{f_{h_1 h_2}'}{n_{h_1 h_2}^0} \left\{ (\mathbf{v}_{h_1 h_2}^{(i)0})^2 - \frac{3k}{m_1} T_{h_1 h_2}^{(i)0} \right\} d\xi_{h_1 h_2}, \\ K_{h_1 h_2}^{(ik)0} &= m_1 \int \frac{f_{h_1 h_2}^0}{n_{h_1 h_2}^0} \mathbf{v}_{h_1 h_2}^{(k)0} \mathbf{v}_{h_1 h_2}^{(i)0} d\xi_{h_1 h_2}, & K_{h_1 h_2}^{(ik)'} &= m_1 \int f_{h_1 h_2}' \mathbf{v}_{h_1 h_2}^{(k)0} \mathbf{v}_{h_1 h_2}^{(i)0} d\xi_{h_1 h_2}, \\ \mathbf{m}_{h_1 h_2}^{(ki)0} &= m_1 \int f_{h_1 h_2}^0 \mathbf{v}_{h_1 h_2}^{(k)0} \frac{1}{2} (\mathbf{v}_{h_1 h_2}^{(i)0})^2 d\xi_{h_1 h_2}, \\ \mathbf{m}_{h_1 h_2}^{(ki)'} &= m_1 \int f_{h_1 h_2}' \left\{ \mathbf{v}_{h_1 h_2}^{(k)0} \left(\frac{1}{2} (\mathbf{v}_{h_1 h_2}^{(i)0})^2 - \frac{3k}{2m_1} T_{h_1 h_2}^{(i)0} \right) - \frac{\mathbf{v}_{h_1 h_2}^{(k)0} \cdot \mathbf{K}_{h_1 h_2}^{(ki)0}}{n_{h_1 h_2}^0 m_1} \right\} d\xi_{h_1 h_2}, \end{aligned} \right\} \quad (50)$$

with similar expressions for $\mathbf{u}_{h_1 h_2}^{(j)}$, $T_{h_1 h_2}^{(j)}$, etc.

It may be mentioned here that since $F_{h_1 h_2}^0$, $n_{h_1 h_2}$, etc., depend on λ_i at the mass centre of the set h_1, h_2 , whenever they are integrated over the co-ordinates of one or more of the molecules in the set h_1, h_2 , they should be expanded inside the integral with respect to the new mass centre of the reduced set, e.g.

$$\int \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} n_{h_1+1, h_2} d\mathbf{x}_1^{(h_1+1)} = \int \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} \left\{ n_{h_1+1, h_2} \right\} + \mathbf{d}_1 \cdot \sum_i \frac{\partial \lambda_i}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_i} \left\{ n_{h_1+1, h_2} \right\} + \dots \Bigg\} d\mathbf{x}_1^{(h_1+1)}, \quad (51)$$

where $\mathbf{d}_1 = \frac{m_1}{M+m_1} (\mathbf{x}_1^{(h_1+1)} - \mathbf{x})$, $M = h_1 m_1 + h_2 m_2$, (52)

the curly bracket $\{ \}$ is used to indicate that quantities inside it depend on λ_i at the new mass centre after integration.

With these remarks about the method of expansion, we can start to expand equation (9). The equation of motion for $F_{h_1 h_2}^0$ is

$$\begin{aligned} [K(h_1 h_2) + W(h_1 h_2), F_{h_1 h_2}^0] + \iint [V_{h_1 h_2}^{(h_1+1)}, \{F_{h_1+1, h_2}^0\}] d\mathbf{x}_1^{(h_1+1)} d\boldsymbol{\xi}_1^{(h_1+1)} \\ + \iint [V_{h_1 h_2}^{(h_2+1)}, \{F_{h_1, h_2+1}^0\}] d\mathbf{x}_2^{(h_2+1)} d\boldsymbol{\xi}_2^{(h_2+1)} = 0, \end{aligned} \quad (53)$$

since the $(\partial \lambda_i / \partial t)$'s in (44) involve no terms not containing $\partial \lambda_i / \partial \mathbf{x}$. The term $[K(h_1 h_2), F_{h_1 h_2}^0]$ in (53) reduces to $-\left(\sum_i \mathbf{v}_1^{(i)} \cdot \frac{\partial}{\partial \mathbf{x}_1^{(i)}} + \sum_j \mathbf{v}_2^{(j)} \cdot \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \right) F_{h_1 h_2}^0$ by virtue of (48). (53) is then seen to be satisfied by the homogeneous equilibrium solution known from statistical mechanics

$$F_{h_1 h_2}^0 = N_{h_1 h_2}^0 \left(\frac{m_1}{2\pi kT} \right)^{\frac{3}{2}h_1} \left(\frac{m_2}{2\pi kT} \right)^{\frac{3}{2}h_2} \exp - \left\{ \frac{m_1}{2kT} \sum_i v_1^{(i)2} + \frac{m_2}{2kT} \sum_j v_2^{(j)2} \right\}. \quad (54)$$

Substituting (54) in (53), one obtains the typical equation after separating equations vanishing independently

$$\begin{aligned} \frac{kT}{N_{h_1 h_2}^0} \frac{\partial N_{h_1 h_2}^0}{\partial \mathbf{x}_1^{(i)}} + \frac{\partial W(h_1 h_2)}{\partial \mathbf{x}_1^{(i)}} + \int \frac{\partial \phi_1^{(i, h_1+1)}}{\partial \mathbf{x}_1^{(i)}} \frac{\{N_{h_1+1, h_2}^0\}}{N_{h_1 h_2}^0} d\mathbf{x}_1^{(h_1+1)} \\ + \int \frac{\partial \psi^{(i, h_2+1)}}{\partial \mathbf{x}_1^{(i)}} \frac{\{N_{h_1, h_2+1}^0\}}{N_{h_1 h_2}^0} d\mathbf{x}_2^{(h_2+1)} = 0, \end{aligned} \quad (55)$$

which is satisfied by

$$N_{h_1 h_2}^0 = \frac{1}{Q_0 (N_1 - h_1)! (N_2 - h_2)!} \int \exp \left[- \frac{W(N_1 N_2)}{kT} \right] d\mathbf{x}_1^{(h_1+1)} \dots d\mathbf{x}_1^{(N_1)} d\mathbf{x}_2^{(h_2+1)} \dots d\mathbf{x}_2^{(N_2)}, \quad (56)$$

where $Q_0 = \frac{1}{N_1! N_2!} \int \exp \left[- \frac{1}{kT} W(N_1 N_2) \right] d\mathbf{x}_{N_1 N_2}. \quad (57)$

One sees here that $N_{h_1 h_2}^0$ can only depend on $\mathbf{r}_1^{(ik)}, \mathbf{r}_2^{(jl)}, \mathbf{s}^{(ij)}, T, n_1$ and n_2 but not on \mathbf{u} , as a constant motion of the whole assembly can obviously have no effect on $N_{h_1 h_2}^0$.

By means of (50), one obtains

$$\left. \begin{aligned} \mathbf{u}_{h_1 h_2}^{(i)0} &= \mathbf{u} = \mathbf{u}_{h_1 h_2}^{(j)0}, & \mathbf{v}_{h_1 h_2}^{(i)0} &= \mathbf{v}_1^{(i)}, & T_{h_1 h_2}^{(i)0} &= T = T_{h_1 h_2}^{(j)0}, \\ K_{h_1 h_2}^{(ji)0} &= N_{h_1 h_2}^0 kT \delta_{ik} 1, & m_{h_1 h_2}^{(ji)0} &= 0, & \text{etc.} \end{aligned} \right\} \quad (58)$$

In particular, for $h_1 = 1, h_2 = 0$ or $h_1 = 0, h_2 = 1$, one has

$$\left. \begin{aligned} n_1^0 &= n_1, \quad \mathbf{u}_1^0 = \mathbf{u} = \mathbf{u}_2^0, \quad T_1^0 = T = T_2^0, \\ K_1^0 &= n_1 k T \mathbf{1}, \quad \mathbf{m}_1^0 = 0 = \mathbf{m}_2^0, \quad \text{etc.} \end{aligned} \right\} \quad (59)$$

The time derivative $\left(\frac{\partial}{\partial t} \lambda_i\right)'$ may now be calculated:

$$\left. \begin{aligned} \left(\frac{\partial}{\partial t} n_1\right)' &= -\frac{\partial}{\partial \mathbf{x}} \cdot (n_1 \mathbf{u}), \\ \left(\frac{\partial}{\partial t} \mathbf{u}\right)' &= -\mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} \mathbf{u} - \frac{1}{\rho} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{K}^0 + \frac{1}{\rho} (\rho_1 \boldsymbol{\eta}'_1 + \rho_2 \boldsymbol{\eta}'_2) = -\mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} \mathbf{u} - \frac{1}{\rho} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{P}^0 + \frac{1}{\rho} \mathbf{F}, \\ \left(\frac{\partial}{\partial t} T\right)' &= -\mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} T - \frac{2}{3kn} \left(\mathbf{K}^0 : \frac{\partial}{\partial \mathbf{x}} \mathbf{u} + \delta'_1 + \delta'_2 \right) = -\mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} T - \frac{2}{3} T (1 - \sigma) \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u}; \end{aligned} \right\} \quad (60)$$

$$\mathbf{P}^0 = \mathbf{K}^0 + \mathbf{L}^0,$$

$$\mathbf{K}^0 = \mathbf{K}_1^0 + \mathbf{K}_2^0 = n k T \mathbf{1},$$

$$\mathbf{L}^0 = \mathbf{L}_1^0 + \mathbf{L}_2^0,$$

$$\mathbf{L}_1^0 = -\frac{1}{2} \int \mathbf{r}_1 \frac{\partial \phi_1}{\partial \mathbf{r}_1} \{N_{20}^0\} d\mathbf{r} - \frac{m_2}{m_1 + m_2} \int \mathbf{s}_1 \frac{\partial \psi}{\partial \mathbf{s}_1} \{N_{11}^0\} d\mathbf{s}_1, \quad \mathbf{r}_1 = \mathbf{r}_1^{(12)}, \quad \mathbf{s}_1 = \mathbf{x}_2^{(1)} - \mathbf{x}_1^{(1)},$$

$$\rho_1 \boldsymbol{\eta}'_1 = n_1 \mathbf{F}_1 - \left\{ \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}_1^0 - \int \{N'_{11}\} \frac{\partial \psi}{\partial \mathbf{s}_1} d\mathbf{s}_1 \right\},$$

$$\rho_1 \boldsymbol{\eta}'_1 + \rho_2 \boldsymbol{\eta}'_2 = \mathbf{F} - \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{L}^0, \quad \mathbf{F} = n_1 \mathbf{F}_1 + n_2 \mathbf{F}_2,$$

$$\delta'_1 + \delta'_2 = -\sigma n k T \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u}, \quad \sigma = \frac{\left(n_1 \frac{\partial}{\partial n_1} + n_2 \frac{\partial}{\partial n_2} + \frac{2}{3} T \frac{\partial}{\partial T} - 1 \right) I}{\left(n k T + \frac{2}{3} T \frac{\partial}{\partial T} I \right)},$$

$$I = \frac{1}{2} \left\{ n_1 \left(\int \frac{\{N_{20}^0\}}{n_1} \phi_1 d\mathbf{x}_1^{(2)} + \int \frac{\{N_{11}^0\}}{n_1} \psi d\mathbf{x}_2^{(1)} \right) + n_2 \left(\int \frac{\{N_{02}^0\}}{n_2} \phi_2 d\mathbf{x}_2^{(2)} + \int \frac{\{N_{11}^0\}}{n_2} \psi d\mathbf{x}_1^{(1)} \right) \right\}.$$

The last three equations in (61) will be proved presently. It has to be pointed out that \mathbf{L}_1^0 is not the whole of the partial potential pressure to the zeroth order but only part of it, the divergence of the rest, which is chiefly responsible for mutual diffusion in non-uniform fluids, is $\int \{N'_{11}\} \frac{\partial \psi}{\partial \mathbf{s}} d\mathbf{s}$. This does not vanish because N'_{11} is not symmetric with respect to the different molecules concerned. The sum $(N'_{11} + N'_{11})$ is, however, symmetric with respect to the two molecules concerned, hence

$$\int \{N'_{11} + N'_{11}\} \frac{\partial \psi}{\partial \mathbf{s}} d\mathbf{s} = 0$$

and \mathbf{L}^0 does not contain terms of this type.

With the help of (60), one can investigate

$$\left(\frac{\partial}{\partial t} n_{h_1 h_2}\right)^0, \quad \left(\frac{\partial}{\partial t} \mathbf{u}_{h_1 h_2}^{(i)}\right)^0 \quad \text{and} \quad \left(\frac{\partial}{\partial t} T_{h_1 h_2}^{(i)}\right)^0$$

to the zeroth order by expanding (13), (24) and (30). They are all found to vanish.

The expansion of $\left(\frac{\partial}{\partial t} \mathbf{u}_{h_1 h_2}^{(i)}\right)^0$ leads to (55). The expansion of $\left(\frac{\partial}{\partial t} n_{h_1 h_2}\right)'$ may be used to obtain the last three expressions in (61), thus:

$$\begin{aligned} \left(\frac{\partial}{\partial t} n_{h_1 h_2}\right)' &= -\frac{\partial}{\partial \mathbf{x}} \cdot (n_1 \mathbf{u}) \frac{\partial N_{h_1 h_2}^0}{\partial n_1} - \frac{\partial}{\partial \mathbf{x}} \cdot (n_2 \mathbf{u}) \frac{\partial N_{h_1 h_2}^0}{\partial n_2} \\ &\quad + \left\{ \mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} T - \frac{2}{3kn} \left(nkT \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \delta'_1 + \delta'_2 \right) \right\} \frac{\partial N_{h_1 h_2}^0}{\partial T}, \\ &\quad \left\{ \sum_i \frac{\partial}{\partial \mathbf{x}_1^{(i)}} \cdot (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(i)}) + \sum_j \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \cdot (n_{h_1 h_2} \mathbf{u}_{h_1 h_2}^{(j)}) \right\}' \\ &= N_{h_1 h_2}^0 \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \sum_i \frac{\partial}{\partial \mathbf{x}_1^{(i)}} \cdot (N_{h_1 h_2}^0 \mathbf{U}_{h_1 h_2}^{(i)'}) + \sum_j \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \cdot (N_{h_1 h_2}^0 \mathbf{U}_{h_1 h_2}^{(j)'}) \\ &\quad + \mathbf{u} \cdot \left(\frac{\partial n_1}{\partial \mathbf{x}} \frac{\partial}{\partial n_1} + \frac{\partial n_2}{\partial \mathbf{x}} \frac{\partial}{\partial n_2} + \frac{\partial T}{\partial \mathbf{x}} \frac{\partial}{\partial T} \right) N_{h_1 h_2}^0. \end{aligned}$$

Combining the above two equations, one finds

$$\begin{aligned} \left(1 - \left(n_1 \frac{\partial}{\partial n_1} + n_2 \frac{\partial}{\partial n_2} + \frac{2}{3} T \frac{\partial}{\partial T} \right) \right) N_{h_1 h_2}^0 \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} - \frac{2}{3kn} (\delta'_1 + \delta'_2) \frac{\partial N_{h_1 h_2}^0}{\partial T} \\ + \sum_i \frac{\partial}{\partial \mathbf{x}_1^{(i)}} \cdot (N_{h_1 h_2}^0 \mathbf{U}_{h_1 h_2}^{(i)'}) + \sum_j \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \cdot (N_{h_1 h_2}^0 \mathbf{U}_{h_1 h_2}^{(j)'}) = 0. \end{aligned} \quad (62)$$

For $h_1 = 2, h_2 = 0$, (62) becomes

$$\frac{\partial}{\partial \mathbf{r}_1} \cdot \{ N_{20}^0 (\dot{\mathbf{U}}_{20}^{(2)'} - \mathbf{U}_{20}^{(1)'}) \} = \Gamma N_{20}^0 \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} - \frac{2}{3kn} (\delta'_1 + \delta'_2) \frac{\partial N_{20}^0}{\partial T}, \quad (63)$$

where

$$\Gamma = \left(n_1 \frac{\partial}{\partial n_1} + n_2 \frac{\partial}{\partial n_2} + \frac{2}{3} T \frac{\partial}{\partial T} - 1 \right). \quad (64)$$

Multiplying (63) by $\frac{1}{2} \phi_1$ and integrating over \mathbf{r}_1 , one obtains

$$\begin{aligned} \frac{1}{2} \int \phi_1 \frac{\partial}{\partial \mathbf{r}_1} \cdot \{ N_{20}^0 (\mathbf{U}_{20}^{(2)'} - \mathbf{U}_{20}^{(1)'}) \} d\mathbf{r}_1 \\ = -\frac{1}{2} \int \frac{\partial \phi_1}{\partial \mathbf{r}_1} \cdot \{ N_{20}^0 (\mathbf{U}_{20}^{(2)'} - \mathbf{U}_{20}^{(1)'}) \} d\mathbf{r}_1 = - \int \frac{\partial \phi_1}{\partial \mathbf{x}_1^{(1)}} \cdot \{ N_{20}^0 (\mathbf{U}_{20}^{(1)'} - \mathbf{u}) \} d\mathbf{x}_1^{(2)} \\ = \Gamma I_1^{(1)} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{2}{3kn} (\delta'_1 + \delta'_2) \frac{\partial I_1^{(1)}}{\partial T}, \end{aligned} \quad (70)$$

where

$$I_1^{(1)} = \frac{1}{2} \int \{ N_{20}^0 \} \phi_1 d\mathbf{x}_1^{(2)}.$$

Similarly, it can be shown that

$$\left. \begin{aligned} - \int \frac{\partial \phi_2}{\partial \mathbf{x}_2^{(1)}} \cdot \{ (\mathbf{U}_{02}^{(1)} - \mathbf{u}) N_{02}^0 \} d\mathbf{x}_2^{(2)} &= \Gamma I_2^{(2)} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{2}{3kn} (\delta'_1 + \delta'_2) \frac{\partial}{\partial T} I_2^{(2)}, \\ - \int \frac{\partial \psi}{\partial \mathbf{x}_1^{(1)}} \cdot \{ (\mathbf{U}_{11}^{(1)} - \mathbf{u}) N_{11}^0 \} d\mathbf{x}_2^{(1)} &= \Gamma I_1^{(2)} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{2}{3kn} (\delta'_1 + \delta'_2) \frac{\partial}{\partial T} I_1^{(2)}, \\ - \int \frac{\partial \psi}{\partial \mathbf{x}_2^{(1)}} \cdot \{ (\mathbf{U}_{11}^{(1)} - \mathbf{u}) N_{11}^0 \} d\mathbf{x}_1^{(1)} &= \Gamma I_2^{(1)} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{2}{3kn} (\delta'_1 + \delta'_2) \frac{\partial}{\partial T} I_2^{(1)}, \end{aligned} \right\} \quad (71)$$

where

$$I_1^{(2)} = \frac{1}{2} \int \{N_{11}^0\} \psi d\mathbf{x}_1^{(2)}, \quad I_2^{(1)} = \frac{1}{2} \int \{N_{11}^0\} \psi d\mathbf{x}_1^{(1)}, \quad I_2^{(2)} = \frac{1}{2} \int \{N_{02}^0\} \phi_2 d\mathbf{x}_2^{(2)}.$$

Adding (70) and (71) together, one obtains

$$-(\delta'_1 + \delta'_2) = \Gamma I \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{2}{3kn} (\delta'_1 + \delta'_2) \frac{\partial}{\partial T} I.$$

Hence it follows that $(\delta'_1 + \delta'_2) = -\sigma nkT \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u}$,

where $\sigma = \frac{\Gamma I}{\left(nkT + \frac{2}{3}T \frac{\partial}{\partial T} I\right)}$, $I = I_1^{(1)} + I_1^{(2)} + I_2^{(1)} + I_2^{(2)}$.

The last three relations in (61) are thus proved.

To obtain the equation of motion for $F'_{h_1 h_2}$, we first write down the first-order terms of (9) after expansion:

$$\begin{aligned} \left(\frac{\partial}{\partial t} f_{h_1 h_2}\right)' &= -\left(\frac{\partial}{\partial \mathbf{x}} \cdot (n_1 \mathbf{u}) \frac{\partial N_{h_1 h_2}^0}{\partial n_1} + \frac{\partial}{\partial \mathbf{x}} \cdot (n_2 \mathbf{u}) \frac{\partial N_{h_1 h_2}^0}{\partial n_2}\right) \frac{F_{h_1 h_2}^0}{N_{h_1 h_2}^0} \\ &\quad + \left\{-\mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} \mathbf{u} - \frac{1}{\rho} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{P}^0 + \frac{\mathbf{F}}{\rho}\right\} \cdot \frac{\mathbf{G}}{kT} F_{h_1 h_2}^0 \\ &\quad + \left(-\mathbf{u} \cdot \frac{\partial}{\partial \mathbf{x}} T - \frac{2}{3}(1-\sigma) T \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u}\right) \left(\frac{T}{N_{h_1 h_2}^0} \frac{\partial N_{h_1 h_2}^0}{\partial T} + \frac{Q}{kT} - \frac{3}{2}h\right) \frac{F_{h_1 h_2}^0}{T}, \\ -[K(h_1 h_2), f_{h_1 h_2}]' &= \left(\mathbf{u} + \frac{\mathbf{G}}{M}\right) \cdot \left\{\left(\frac{\partial n_1}{\partial \mathbf{x}} \frac{\partial N_{h_1 h_2}^0}{\partial n_1} + \frac{\partial n_2}{\partial \mathbf{x}} \frac{\partial N_{h_1 h_2}^0}{\partial n_2}\right) \frac{F_{h_1 h_2}^0}{N_{h_1 h_2}^0}\right. \\ &\quad \left.+ \frac{\partial}{\partial \mathbf{x}} \mathbf{u} \cdot \mathbf{G} \frac{F_{h_1 h_2}^0}{kT} + \frac{\partial T}{\partial \mathbf{x}} \left(\frac{T}{N_{h_1 h_2}^0} \frac{\partial N_{h_1 h_2}^0}{\partial T} + \frac{Q}{kT} - \frac{3}{2}h\right) \frac{F_{h_1 h_2}^0}{T}\right\} \\ &\quad - [K(h_1 h_2), F'_{h_1 h_2}], \end{aligned}$$

$$[W(h_1 h_2) + E(h_1 h_2), f_{h_1 h_2}]' = [W(h_1 h_2), F'_{h_1 h_2}] + \left(\sum_i \mathbf{F}_i^{(i)} \cdot \mathbf{v}_1^{(i)} + \sum_j \mathbf{F}_j^{(j)} \cdot \mathbf{v}_2^{(j)}\right) \frac{F_{h_1 h_2}^0}{kT},$$

$$\begin{aligned} &\left(\iint [V_{h_1 h_2}^{(h_1+1)}, f_{h_1+1, h_2}] d\mathbf{x}_1^{(h_1+1)} d\xi_{51}^{(h_1+1)}\right)' \\ &= \iint \left[V_{h_1 h_2}^{(h_1+1)}, F'_{h_1+1, h_2} + \mathbf{d}_1 \cdot \sum_i \left(\frac{\partial \lambda_i}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_i} F_{h_1+1, h_2}^0\right)\right] d\mathbf{x}_1^{(h_1+1)} d\xi_{51}^{(h_1+1)}, \\ &\left(\iint [V_{h_1 h_2}^{(h_2+1)}, f_{h_1, h_2+1}] d\mathbf{x}_2^{(h_2+1)} d\xi_{52}^{(h_2+1)}\right)' \\ &= \iint \left[V_{h_1 h_2}^{(h_2+1)}, F'_{h_1, h_2+1} + \mathbf{d}_2 \cdot \sum_i \left(\frac{\partial \lambda_i}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_i} F_{h_1, h_2+1}^0\right)\right] d\mathbf{x}_2^{(h_2+1)} d\xi_{52}^{(h_2+1)}, \end{aligned}$$

where

$$\mathbf{G} = \sum_i m_1 \mathbf{v}_1^{(i)} + \sum_j m_2 \mathbf{v}_2^{(j)}, \quad \mathbf{d}_1 = \frac{m_1}{M + m_1} (\mathbf{x}_1^{(h_1+1)} - \mathbf{x}),$$

$$Q = \frac{1}{2} (m_1 \sum_i v_1^{(i)2} + m_2 \sum_j v_2^{(j)2}), \quad \mathbf{d}_2 = \frac{m_2}{M + m_2} (\mathbf{x}_2^{(h_2+1)} - \mathbf{x}),$$

$$M = h_1 m_1 + h_2 m_2,$$

$$h = h_1 + m_2.$$

Combining the above results and rearranging terms, one obtains

$$\begin{aligned}
 & F_{h_1 h_2}^0 \mathbf{G} \cdot \left\{ \frac{1}{MN_{h_1 h_2}^0} \left(\frac{\partial n_1}{\partial \mathbf{x}} \frac{\partial}{\partial n_1} + \frac{\partial n_2}{\partial \mathbf{x}} \frac{\partial}{\partial n_2} \right) N_{h_1 h_2}^0 - \frac{1}{\rho} \frac{\partial n}{\partial \mathbf{x}} - \frac{1}{\rho k T} \left(\frac{\partial n_1}{\partial \mathbf{x}} \frac{\partial}{\partial n_1} + \frac{\partial n_2}{\partial \mathbf{x}} \frac{\partial}{\partial n_2} \right) L^0 \right\} \\
 & - F_{h_1 h_2}^0 \left\{ \frac{1}{N_{h_1 h_2}^0} \mathcal{D} N_{h_1 h_2}^0 + \frac{2}{3} (1 - \sigma) \left(\frac{T}{N_{h_1 h_2}^0} \frac{\partial N_{h_1 h_2}^0}{\partial T} + \frac{Q}{k T} - \frac{3}{2} h \right) - \frac{1}{3} \frac{G^2}{M k T} \right\} \\
 & \times \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{F_{h_1 h_2}^0}{M k T} \mathbf{G}^0 \mathbf{G} : \mathbf{b} + \frac{F_{h_1 h_2}^0}{T} \left\{ \frac{T}{N_{h_1 h_2}^0} \frac{\partial N_{h_1 h_2}^0}{\partial T} + \frac{Q}{k T} - \frac{3}{2} h - \frac{M n}{\rho} - \frac{M}{k \rho} \frac{\partial L^0}{\partial T} \right\} \\
 & \times \frac{\mathbf{G}}{M} \cdot \frac{\partial T}{\partial \mathbf{x}} + F_{h_1 h_2}^0 \left\{ \frac{\mathbf{E} \cdot \mathbf{G}}{\rho k T} + \frac{\mathbf{G}_1 \cdot \mathbf{F}_1}{m_1 k T} + \frac{\mathbf{G}_2 \cdot \mathbf{F}_2}{m_2 k T} \right\} + \left(\sum_i \mathbf{v}_1^{(i)} \cdot \frac{\partial}{\partial \mathbf{x}_1^{(i)}} + \sum_j \mathbf{v}_2^{(j)} \cdot \frac{\partial}{\partial \mathbf{x}_2^{(j)}} \right) F'_{h_1 h_2} \\
 & = [W(h_1 h_2), F'_{h_1 h_2}] + \iint \left[V_{h_1 h_2}^{(h_1+1)}, \left\{ F'_{h_1+1, h_2} + \mathbf{d}_1 \cdot \sum_l \frac{\partial \lambda_l}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_l} F_{h_1+1, h_2}^0 \right\} \right] d\mathbf{x}_1^{(h_1+1)} d\xi_1^{(h_1+1)} \\
 & + \iint \left[V_{h_1 h_2}^{(h_2+1)}, \left\{ F'_{h_1, h_2+1} + \mathbf{d}_2 \cdot \sum_l \frac{\partial \lambda_l}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_l} F_{h_1, h_2+1}^0 \right\} \right] d\mathbf{x}_2^{(h_2+1)} d\xi_2^{(h_2+1)}, \quad (72)
 \end{aligned}$$

where $\mathbf{G}_1 = \sum_i m_1 \mathbf{v}_1^{(i)}$, $\mathbf{G}_2 = \sum_j m_2 \mathbf{v}_2^{(j)}$, $\mathcal{D} = n_1 \frac{\partial}{\partial n_1} + n_2 \frac{\partial}{\partial n_2}$, and \mathbf{b} is the symmetrical, non-divergent part of $\frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u}$. For $h_1 = 1$, $h_2 = 0$, equation (72) becomes

$$\begin{aligned}
 & F_1^0 \left\{ \frac{1}{\rho_1} \frac{\partial n_1}{\partial \mathbf{x}} - \frac{1}{\rho} \frac{\partial n}{\partial \mathbf{x}} - \frac{1}{\rho k T} \left(\frac{\partial n_1}{\partial \mathbf{x}} \frac{\partial}{\partial n_1} + \frac{\partial n_2}{\partial \mathbf{x}} \frac{\partial}{\partial n_2} \right) L^0 \right\} \cdot m_1 \mathbf{v}_1 + F_1^0 \frac{m_1}{k T} \mathbf{v}_1 \mathbf{v}_1 : \mathbf{b} \\
 & + F_1^0 \left\{ \sigma \left(\frac{m_1 v_1^2}{3 k T} - 1 \right) \right\} \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{F_1^0}{T} \left\{ \frac{m_1 v_1^2}{2 k T} - \frac{5}{2} + \frac{\rho - n m_1}{\rho} - \frac{m_1}{k \rho} \frac{\partial L^0}{\partial T} \right\} \frac{\partial T}{\partial \mathbf{x}} \cdot \mathbf{v}_1 + \frac{F_1^0}{k T} \left(\frac{m \mathbf{F}_1}{\rho} - \mathbf{F}_2 \right) \cdot \mathbf{v}_1 \\
 & = \frac{1}{m_1} \iint \frac{\partial \phi_1}{\partial \mathbf{x}_1^{(1)}} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \left\{ F'_{20} + \frac{I_1}{2} \cdot \sum_l \frac{\partial \lambda_l}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_l} F_{20}^0 \right\} d\mathbf{x}_1^{(2)} d\xi_1^{(2)} \\
 & + \frac{1}{m_2} \iint \frac{\partial \psi}{\partial \mathbf{x}_1^{(1)}} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \left\{ F'_{11} + \frac{\mathbf{s}_1}{2} \cdot \sum_l \frac{\partial \lambda_l}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_l} F_{11}^0 \right\} d\mathbf{x}_2^{(1)} d\xi_2^{(1)}. \quad (73)
 \end{aligned}$$

The right-hand side of the above equation after substituting the known solutions of F_{20}^0 and F_{11}^0 , and performing integration, becomes

$$\begin{aligned}
 & - \frac{1}{m_1} \iint \frac{\partial \phi_1}{\partial \mathbf{x}_1^{(1)}} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \{ F'_{20} \} \partial \mathbf{x}_1^{(2)} \partial \xi_1^{(2)} - \frac{1}{m_2} \iint \frac{\partial \psi}{\partial \mathbf{x}_1^{(1)}} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \{ F'_{11} \} d\mathbf{x}_2^{(1)} d\xi_2^{(1)} \\
 & - F_1^0 \frac{\mathbf{v}_1}{n_1 k T} \cdot \sum_l \left(\frac{\partial \lambda_l}{\partial \mathbf{x}} \frac{\partial}{\partial \lambda_l} \right) L_1^0 - \frac{F_1^0}{T} \frac{L_1^0}{n_1 k T} \left(\frac{m_1 v_1^2}{2 k T} - \frac{5}{2} \right) \frac{\partial T}{\partial \mathbf{x}} \cdot \mathbf{v}_1 \\
 & - F_1^0 \frac{L_1^0}{n_1 k T} \left(\frac{m_1}{k T} \right) \mathbf{v}_1 \mathbf{v}_1 : \mathbf{b} - F_1^0 \frac{L_1^0}{n_1 k T} \left(\frac{m_1 v_1^2}{3 k T} - 1 \right) \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u}. \quad (74)
 \end{aligned}$$

Combining (73) and (74), one has

$$\begin{aligned}
 & \frac{F_1^0 \rho_1 \rho_2}{n_1 \rho} \left\{ \frac{1}{\rho_1} + \frac{1}{kT} \left(\frac{1}{\rho_1} \frac{\partial L_1^0}{\partial n_1} - \frac{1}{\rho_2} \frac{\partial L_2^0}{\partial n} \right) \right\} \mathbf{v}_1 \cdot \frac{\partial n_1}{\partial \mathbf{x}} - \frac{F_1^0 \rho_1 \rho_2}{n_1 \rho} \left\{ \frac{1}{\rho_2} + \frac{1}{kT} \left(\frac{1}{\rho_2} \frac{\partial L_2^0}{\partial n_2} - \frac{1}{\rho_1} \frac{\partial L_1^0}{\partial n_2} \right) \right\} \mathbf{v}_1 \cdot \frac{\partial n_2}{\partial \mathbf{x}} \\
 & + \frac{F_1^0}{T} \left\{ \left(1 + \frac{L_1^0}{n_1 kT} \right) \left(\frac{m_1 v_1^2}{2kT} - \frac{5}{2} \right) + \frac{1}{n_1} \left[\frac{n_1 n_2}{\rho} (m_2 - m_1) + \frac{\rho_1 \rho_2}{k\rho} \left(\frac{1}{\rho_1} \frac{\partial L_1^0}{\partial T} - \frac{1}{\rho_2} \frac{\partial L_2^0}{\partial T} \right) \right] \right\} \mathbf{v}_1 \cdot \frac{\partial T}{\partial \mathbf{x}} \\
 & + F_1^0 \frac{m_1}{kT} \left(1 + \frac{L_1^0}{n_1 kT} \right) \mathbf{v}_1 \mathbf{v}_1 : \mathbf{b} + F_1^0 \left(\sigma_1 + \frac{L_1^0}{n_1 kT} \right) \left(\frac{m_1 v_1^2}{3kT} - 1 \right) \frac{\partial}{\partial \mathbf{x}} \cdot \mathbf{u} + \frac{F_1^0 \rho_1 \rho_2}{n_1 \rho kT} \mathbf{a} \cdot \mathbf{v}_1 \\
 & = \frac{-1}{m_1} \left\{ \iint \frac{\partial \phi_1}{\partial \mathbf{r}_1} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \{ F'_{20} \} d\mathbf{r}_1 d\xi_1^{(2)} + \iint \frac{\partial \psi}{\partial \mathbf{s}_1} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \{ F'_{11} \} d\mathbf{s}_1 d\xi_2^{(1)} \right\}, \quad (75)
 \end{aligned}$$

where $\mathbf{a} = \frac{\mathbf{F}_2}{m_2} - \frac{\mathbf{F}_1}{m_1}$ is the difference of acceleration of the two kinds of molecules in the external field.

A similar equation can by symmetry be written down for F'_{02} , F'_{11} and F_2^0 . Equation (75) is an exact analogue of the corresponding equation in gas theory (cf. Chapman & Cowling 1939). Though (75) is successful in expressing the rate of change of f_1 in terms of f'_{20} and f'_{11} corresponding to the collision integral in the gas theory, yet in contrast to the latter where the statistical effect of the predominant mechanism of interaction (binary encounters) has been explicitly taken into account, it has only formal value as it stands. However, there are two possible ways of solving the problem. The first is to proceed to equations containing more than two molecules in the set h_1, h_2 and use approximations similar to Kirkwood's approximation to the triple distribution function in terms of the pair distribution functions, thus obtaining closed equations; but the complications that arise for a mixture are likely to be considerable. The second and much more practical way of approach consists in modifying the binary encounter in a manner to be explained in the following section.

6. DISCUSSION OF THE METHOD OF SOLUTION

The assumption of binary encounters to which the whole theory of gases owes its extensive development loses its validity as the density increases. In the liquid region, more refined considerations than the binary encounter assumption have to be introduced; as the consideration of multiple encounter is impracticable if not impossible, we propose here to modify the binary encounter assumption so as to take into account the influence of the presence of the other molecules. We assume that it is possible to account for the average relative trajectory of a pair of molecules in a liquid during a period of intense interaction with each other by means of an apparent potential between the pair. This assumption is suggested in the equilibrium state by the fact that the equilibrium pair distribution function f_{20}^0 can be written in the form

$$f_{20}^0 = n_{20}^0 \frac{f^0}{(n_1)^2} = e^{-\phi_1^*/kT},$$

which serves to define the apparent potential ϕ_1^* . From this equation, one sees that the distribution function n_{20}^0 is attributable to an apparent potential ϕ_1^* between the pair which is independent of their velocities. From (55), we have for $h_1 = 2$, $h_2 = 0$

$$-\frac{kT}{n_{20}^0} \frac{\partial n_{20}^0}{\partial \mathbf{x}_1^{(2)}} = \frac{\partial \phi_1^{(12)}}{\partial \mathbf{x}_1^{(2)}} + \int \frac{n_{30}^0}{n_{20}^0} \frac{\partial \phi_1^{(23)}}{\partial \mathbf{x}_1^{(2)}} d\mathbf{x}_1^{(3)} + \int \frac{n_{21}^0}{n_{20}^0} \frac{\partial \psi}{\partial \mathbf{x}_1^{(2)}} d\mathbf{x}_2^{(1)} = \frac{\partial \phi_1^*}{\partial \mathbf{x}_1^{(2)}},$$

from which ϕ_1^* can be found with the help of Kirkwood's approximation (E. A. Rodriguez 1949).

The trajectory thus calculated with the use of ϕ_1^* would represent the average relative trajectory of the pair in the presence of the other molecules in the equilibrium state for each configuration of approach specified by the impact parameter b and the relative velocity $g = |\xi_1^{(2)} - \xi_1^{(1)}|$. Though in the non-equilibrium state, the distortion of the distribution function may give rise to modifications of the apparent potential ϕ_1^* , we shall, as a first approximation, use the ϕ_1^* defined above to calculate the average trajectory of a pair in a non-equilibrium fluid.

The use of ϕ_1^* enables one to reduce a many-body problem to a two-body problem in an approximate manner. The right-hand side of (75) can now be transformed, the typical integral being of the form

$$-\frac{1}{m_1} \iint \frac{\partial \phi_1}{\partial \mathbf{x}_1} \cdot \frac{\partial}{\partial \xi_1^{(1)}} f_{20} d\xi_1^{(2)} d\mathbf{r}_1.$$

The integration over \mathbf{r}_1 needs only to extend over a sphere of influence of radius R_0 within which ϕ_1^* is appreciable. We shall first express f_{20} in terms of a product of two f_1 's for the two molecules concerned. At time t let the velocity of a molecule at $\mathbf{x}_1^{(1)}$ be $\xi_1^{(1)}$, and that of a molecule at $\mathbf{x}_1^{(2)}$ be $\xi_1^{(2)}$. Using ϕ_1^* , one can calculate the trajectory of a molecule at $\mathbf{x}_1^{(2)}$ within the sphere of influence R_0 relative to the one at $\mathbf{x}_1^{(1)}$ for every given relative velocity $\mathbf{g} = \xi_1^{(2)} - \xi_1^{(1)}$. In particular, one can calculate the position and velocity co-ordinates $\mathbf{x}_1^{(1)'}$, $\mathbf{x}_1^{(2)'}$, $\xi_1^{(1)'}$ and $\xi_1^{(2)'}$ of the pair at time t' , when they last approached within distance R_0 of each other, as functions of $\mathbf{x}_1^{(1)}$, \mathbf{r}_1 , $\xi_1^{(1)}$ and $\xi_1^{(2)}$. (In fact, only the velocities $\xi_1^{(1)'}$ and $\xi_1^{(2)'}$ at t' as functions of $\mathbf{x}_1^{(1)}$, \mathbf{r}_1 , $\xi_1^{(1)}$ and $\xi_1^{(2)}$ are required, as will be seen presently.)

One can then write as the limiting form of f_{20}

$$f_{20}(t, \mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \xi_1^{(1)}, \xi_1^{(2)}) = f_{20}(t', \mathbf{x}_1^{(1)'}, \mathbf{x}_1^{(2)'}, \xi_1^{(1)'}, \xi_1^{(2)'}) = f_1(t', \mathbf{x}_1^{(1)'}, \xi_1^{(1)'}) f_1(t', \mathbf{x}_1^{(2)'}, \xi_1^{(2)'}) .$$

The first equality holds because by using ϕ_1^* , etc., Liouville's theorem may be assumed to hold for the pair, and the second equality follows from an application of the law of probability governing independent events. Each of the f_1 's on the right-hand side of the above equation can be expanded in the manner

$$f_1 = f_1^0 + f_1' + \dots,$$

where f_1' represents the deviation from a Maxwellian distribution. We have then for f_{20} after expansion

$$f_{20}(t, \mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \xi_1^{(1)}, \xi_1^{(2)}) = f_1^0(t', \mathbf{x}_1^{(1)'}, \xi_1^{(1)'}) f_1^0(t', \mathbf{x}_1^{(2)'}, \xi_1^{(2)'}) + f_1^0(t', \mathbf{x}_1^{(1)'}, \xi_1^{(1)'}) f_1'(t', \mathbf{x}_1^{(2)'}, \xi_1^{(2)'}) \\ + f_1'(t', \mathbf{x}_1^{(1)'}, \xi_1^{(1)'}) f_1^0(t', \mathbf{x}_1^{(2)'}, \xi_1^{(2)'}) + \dots$$

The first term of $f_1^0(t', \mathbf{x}_1^{(1)}, \xi_1^{(1)}) f_1^0(t', \mathbf{x}_1^{(2)}, \xi_1^{(2)})$ which equals $f_{20}^0(t, \mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \xi_1^{(1)}, \xi_1^{(2)})$ has already been taken into account in (74) in a different manner. For the factor $\{F'_{20}\}$ in (75), one has then

$$\{F'_{20}\} = F_1^0(\mathbf{v}_1^{(1)}, \lambda_i) F_1'(\mathbf{v}_1^{(2)}, \lambda_i, \frac{\partial \lambda_i}{\partial \mathbf{x}}) + F_1'(\mathbf{v}_1^{(1)}, \lambda_i, \frac{\partial \lambda_i}{\partial \mathbf{x}}) F_1^0(\mathbf{v}_1^{(2)}, \lambda_i), \quad (76)$$

where $\mathbf{v}_1^{(1)} = \xi_1^{(1)} - \mathbf{u}(\mathbf{x}_1^{(1)})$, $\mathbf{v}_1^{(2)} = \xi_1^{(2)} - \mathbf{u}(\mathbf{x}_1^{(1)})$ and the λ_i 's and $\partial \lambda_i / \partial \mathbf{x}$'s always refer to the value at $\mathbf{x}_1^{(1)}$ as indicated by the curly bracket.

The right-hand side of (75) now becomes

$$\begin{aligned} & -\frac{1}{m_1} \int \int \frac{\partial \phi_1}{\partial \mathbf{r}_1} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \left\{ F_1^0(\mathbf{v}_1^{(1)}, \lambda_i) F_1'(\mathbf{v}_1^{(2)}, \lambda_i, \frac{\partial \lambda_i}{\partial \mathbf{x}}) + F_1'(\mathbf{v}_1^{(1)}, \lambda_i, \frac{\partial \lambda_i}{\partial \mathbf{x}}) F_1^0(\mathbf{v}_1^{(2)}, \lambda_i) \right\} d\xi_1^{(2)} d\mathbf{r}_1, \\ & -\frac{1}{m} \int \int \frac{\partial \psi}{\partial \mathbf{s}} \cdot \frac{\partial}{\partial \xi_1^{(1)}} \left\{ F_1^0(\mathbf{v}_1^{(1)}, \lambda_i) F_2'(\mathbf{v}_2^{(1)}, \lambda_i, \frac{\partial \lambda_i}{\partial \mathbf{x}}) + F_1'(\mathbf{v}_1^{(1)}, \lambda_i, \frac{\partial \lambda_i}{\partial \mathbf{x}}) F_2^0(\mathbf{v}_2^{(1)}, \lambda_i) \right\} d\xi_1^{(2)} d\mathbf{s}, \end{aligned} \quad (75')$$

where $\mathbf{v}_1^{(1)} = \xi_1^{(1)} - \mathbf{u}(\mathbf{x}_1^{(1)})$ and $\mathbf{v}_2^{(1)} = \xi_2^{(1)} - \mathbf{u}(\mathbf{x}_1^{(1)})$ are functions of \mathbf{s} , $\xi_1^{(1)}$ and $\xi_2^{(1)}$ calculated from an apparent potential ψ^* between a pair of different molecules in the same manner in which $\xi_1^{(1)}$ and $\xi_1^{(2)}$ are obtained.

A closed integral equation for F'_1 is thus obtained. Instead of solving for F'_1 , one needs only to find its first and second moments in the velocity for the purpose of calculating the coefficients of diffusion and viscosity. The problem now consists in first finding the apparent potentials ϕ_1^* , ψ^* , etc., then calculating the trajectory and expressing $\xi_1^{(1)}$, $\xi_1^{(2)}$ as functions of \mathbf{r}_1 , $\xi_1^{(1)}$, $\xi_1^{(2)}$ and $\xi_1^{(1)}$, $\xi_2^{(1)}$ as functions of \mathbf{s} , $\xi_1^{(1)}$, $\xi_2^{(1)}$, and finally solving the integral equation (75) with its right-hand side replaced by (75'). In this way, one can regard (75) as a generalization to Boltzmann's equation in gas theory. The solution of this integral equation remains to be completed.

7. GENERAL EXPRESSIONS FOR THE COEFFICIENTS OF ORDINARY AND THERMAL DIFFUSION AND OF VISCOSITY IN LIQUID MIXTURES

The quantities required for calculating the coefficients of ordinary and thermal diffusion are F'_1 and F'_2 , those for calculating the coefficient of viscosity are F'_1 , F'_2 , N'_{20} , N'_{11} , etc. From quite general considerations the form of the above quantities can be determined.

First consider F'_1 and F'_2 : they are linear in $\partial \lambda_i / \partial \mathbf{x}$ and satisfy the conditions

$$\left. \begin{aligned} n'_1 &= \int F'_1 d\mathbf{v}_1 = 0, \quad n'_2 = \int F'_2 d\mathbf{v}_2 = 0, \\ \rho \mathbf{u}' &= \rho_1 \mathbf{u}'_1 + \rho_2 \mathbf{u}'_2 = m_1 \int F'_1 \mathbf{v}_1 d\mathbf{v}_1 + m_2 \int F'_2 \mathbf{v}_2 d\mathbf{v}_2 = 0, \\ nT' &= n_1 T'_1 + n_2 T'_2 = \frac{2m_1}{3k} \int F'_1 \left(\frac{1}{2} v_1^2 - \frac{3k}{2m_1} T \right) d\mathbf{v}_1 + \frac{2m_2}{3k} \int F'_2 \left(\frac{1}{2} v_2^2 - \frac{3k}{2m_2} T \right) d\mathbf{v}_2 \\ &= \frac{1}{3k} \left(m_1 \int F'_1 v_1^2 d\mathbf{v}_1 + m_2 \int F'_2 v_2^2 d\mathbf{v}_2 \right) = 0, \end{aligned} \right\} \quad (77)$$

which require that any further correction to the velocity distribution F_1^0 and F_2^0 shall not alter the values of the local parameters λ_i . (For the present purpose, the presence of an external force field will not be considered.) Hence F_1' and F_2' must be of the form

$$\left. \begin{aligned} F_1' &= \varphi_1^{(1)} \mathbf{v}_1 \cdot \frac{\partial n_1}{\partial \mathbf{x}} + \varphi_1^{(2)} \mathbf{v}_1 \cdot \frac{\partial n_2}{\partial \mathbf{x}} + \varphi_1^{(3)} \mathbf{v}_1 \cdot \frac{\partial T}{\partial \mathbf{x}} + \varphi_1^{(4)} \mathbf{v}_1 \mathbf{v}_1 : \mathbf{b}, \\ F_2' &= \varphi_2^{(1)} \mathbf{v}_2 \cdot \frac{\partial n_1}{\partial \mathbf{x}} + \varphi_2^{(2)} \mathbf{v}_2 \cdot \frac{\partial n_2}{\partial \mathbf{x}} + \varphi_2^{(3)} \mathbf{v}_2 \cdot \frac{\partial T}{\partial \mathbf{x}} + \varphi_2^{(4)} \mathbf{v}_2 \mathbf{v}_2 : \mathbf{b}, \end{aligned} \right\} \quad (78)$$

where the φ_i 's are functions of v_i , n_1 , n_2 and T , the φ_2 's functions of v_2 , n_1 , n_2 and T , and

$$m_1 \int \varphi_1^{(i)} v_1^2 d\mathbf{v}_1 + m_2 \int \varphi_2^{(i)} v_2^2 d\mathbf{v}_2 = 0 \quad (i = 1, 2, 3). \quad (79)$$

Next consider N'_{20} : it is linear in $\partial \lambda_i / \partial \mathbf{x}$, even in \mathbf{r}_1 and satisfies $\int N'_{20} d\mathbf{r}_1 = 0$. Hence it must be of the form

$$N'_{20} = \nu_{20}^{(4)} \mathbf{r}_1 \mathbf{r}_1 : \mathbf{b} \quad (80)$$

where $\nu_{20}^{(4)}$ is a function of r_1 , n_1 , n_2 and T . Similarly, one can write down for N'_{02}

$$N'_{02} = \nu_{02}^{(4)} \mathbf{r}_2 \mathbf{r}_2 : \mathbf{b} \quad (81)$$

The forms of N'_{11} and $N'_{\bar{1}\bar{1}}$ are, however, not so simple, since neither of them is even in the relative position vector or symmetrical with respect to the two molecules concerned; but the sum ($N'_{11} + N'_{\bar{1}\bar{1}}$) is an even function of \mathbf{s}_1 ($= \mathbf{x}_2^{(1)} - \mathbf{x}_1^{(1)}$) and the difference ($N'_{11} - N'_{\bar{1}\bar{1}}$) an odd function of \mathbf{s}_1 . Hence

$$\begin{aligned} N'_{11} + N'_{\bar{1}\bar{1}} &= 2\nu_{11}^{(4)} \mathbf{s}_1 \mathbf{s}_1 : \mathbf{b}, \\ N'_{11} - N'_{\bar{1}\bar{1}} &= 2\nu_{11}^{(1)} \mathbf{s}_1 \cdot \frac{\partial n_1}{\partial \mathbf{x}} + 2\nu_{11}^{(2)} \mathbf{s}_1 \cdot \frac{\partial n_2}{\partial \mathbf{x}} + 2\nu_{11}^{(3)} \mathbf{s}_1 \cdot \frac{\partial T}{\partial \mathbf{x}}. \end{aligned}$$

It follows from the above equations that

$$\left. \begin{aligned} N'_{11} &= \nu_{11}^{(1)} \mathbf{s}_1 \cdot \frac{\partial n_1}{\partial \mathbf{x}} + \nu_{11}^{(2)} \mathbf{s}_1 \cdot \frac{\partial n_2}{\partial \mathbf{x}} + \nu_{11}^{(3)} \mathbf{s}_1 \cdot \frac{\partial T}{\partial \mathbf{x}} + \nu_{11}^{(4)} \mathbf{s}_1 \mathbf{s}_1 : \mathbf{b}, \\ N'_{\bar{1}\bar{1}} &= \nu_{11}^{(1)} \mathbf{s}_2 \cdot \frac{\partial n_1}{\partial \mathbf{x}} + \nu_{11}^{(2)} \mathbf{s}_2 \cdot \frac{\partial n_2}{\partial \mathbf{x}} + \nu_{11}^{(3)} \mathbf{s}_2 \cdot \frac{\partial T}{\partial \mathbf{x}} + \nu_{11}^{(4)} \mathbf{s}_2 \mathbf{s}_2 : \mathbf{b}, \end{aligned} \right\} \quad (82)$$

where $\mathbf{s}_2 = -\mathbf{s}_1$.

With these general expressions in hand, it is easy to find the coefficients of diffusion and viscosity in terms of them. Consider diffusion first: the mutual diffusion velocity is

$$\mathbf{u}'_1 - \mathbf{u}'_2 = \int \frac{F'_1}{n_1} \mathbf{v}_1 d\mathbf{v}_1 - \int \frac{F'_2}{n_2} \mathbf{v}_2 d\mathbf{v}_2.$$

As the coefficients of diffusion are measured under constant pressure, it is convenient first to transform the gradients $\partial n_i / \partial \mathbf{x}$, $\partial n_2 / \partial \mathbf{x}$ and $\partial T / \partial \mathbf{x}$ appearing in the formal expressions of F'_1 and F'_2 to $\partial c_i / \partial \mathbf{x}$, $\partial p / \partial \mathbf{x}$ and $\partial T / \partial \mathbf{x}$, according to (19) in part I of the present paper with p given by P^0 in (61). Thus

$$F'_1 = \Omega_1^{(1)} \mathbf{v}_1 \cdot \frac{\partial c_1}{\partial \mathbf{x}} + \Omega_1^{(2)} \mathbf{v}_1 \cdot \frac{\partial p}{\partial \mathbf{x}} + \Omega_1^{(3)} \mathbf{v}_1 \cdot \frac{\partial T}{\partial \mathbf{x}} + \varphi_1^{(4)} \mathbf{v}_1 \mathbf{v}_1 : \mathbf{b}, \quad (83)$$

where

$$\left. \begin{aligned} \Omega_1^{(1)} &= \frac{n^2}{\mathcal{D}p^0} \left(\varphi_1^{(1)} \frac{\partial p^0}{\partial n_2} - \varphi_1^{(2)} \frac{\partial p^0}{\partial n_1} \right), \\ \Omega_1^{(2)} &= \frac{1}{\mathcal{D}p^0} (\varphi_1^{(1)} n_1 + \varphi_1^{(2)} n_2), \\ \Omega_1^{(3)} &= \varphi_1^{(3)} - \frac{\partial p}{\partial T} \frac{(n_1 \varphi_1^{(1)} + n_2 \varphi_1^{(2)})}{\mathcal{D}p^0}, \\ \mathcal{D} &= n_1 \frac{\partial}{\partial n_1} + n_2 \frac{\partial}{\partial n_2}, \\ p^0 &= nkT + \frac{1}{6} \left\{ \int N_{20}^0 \phi_1'(r_1) r_1 d\mathbf{r}_1 + \int N_{02}^0 \phi_2'(r_2) r_2 d\mathbf{r}_2 + 2 \int N_{11}^0 \psi'(s_1) s_1 d\mathbf{s}_1 \right\}. \end{aligned} \right\} \quad (84)$$

A similar transformation is applied to F'_2 . The mutual diffusion velocity ($\mathbf{u}'_1 - \mathbf{u}'_2$) can be expressed as

$$\begin{aligned} \mathbf{u}'_1 - \mathbf{u}'_2 &= \frac{1}{3} \left\{ \int \Omega_1^{(1)} \frac{v_1^2}{n_1} d\mathbf{v}_1 + \int \Omega_2^{(1)} \frac{v_2^2}{n_2} d\mathbf{v}_2 \right\} \frac{\partial c_1}{\partial \mathbf{x}} + \frac{1}{3} \left\{ \int \Omega_1^{(2)} \frac{v_1^2}{n_1} d\mathbf{v}_1 - \int \Omega_2^{(2)} \frac{v_2^2}{n_2} d\mathbf{v}_2 \right\} \frac{\partial p}{\partial \mathbf{x}} \\ &\quad + \frac{1}{3} \left\{ \int \Omega_1^{(3)} \frac{v_1^2}{n_1} d\mathbf{v}_1 - \int \Omega_2^{(3)} \frac{v_2^2}{n_2} d\mathbf{v}_2 \right\} \frac{\partial T}{\partial \mathbf{x}}, \end{aligned} \quad (85)$$

where the term involving $\partial p / \partial \mathbf{x}$ can be dropped under the assumption of constant pressure. Upon comparing (85) with the usual definition of the coefficient of ordinary diffusion D_0 and thermal diffusion D_T

$$\mathbf{u}'_1 - \mathbf{u}'_2 = - \frac{n^2}{n_1 n_2} \left(D_0 \frac{\partial c_1}{\partial \mathbf{x}} - \frac{D_T}{T} \frac{\partial T}{\partial \mathbf{x}} \right),$$

one finds

$$\left. \begin{aligned} D_0 &= - \frac{n_1 n_2}{3n^2} \left\{ \int \Omega_1^{(1)} \frac{v_1^2}{n_1} d\mathbf{v}_1 + \int \Omega_2^{(1)} \frac{v_2^2}{n_2} d\mathbf{v}_2 \right\}, \\ D_T &= - \frac{n_1 n_2}{3n^2} T \left\{ \int \Omega_1^{(3)} \frac{v_1^2}{n_1} d\mathbf{v}_1 - \int \Omega_2^{(3)} \frac{v_2^2}{n_2} d\mathbf{v}_2 \right\}. \end{aligned} \right\} \quad (86)$$

Next consider viscosity: the pressure tensor in the presence of linear gradients $\partial \lambda_i / \partial \mathbf{x}$ is $\mathbf{p} = \mathbf{p}^0 + \mathbf{p}'$, where \mathbf{p}^0 is given by (61) and \mathbf{p}' by

$$\left. \begin{aligned} \mathbf{p}' &= \mathbf{K}' + \mathbf{L}', \\ \mathbf{K}' &= m_1 \int F'_1 \mathbf{v}_1 \mathbf{v}_1 d\mathbf{v}_1 + m_2 \int F'_2 \mathbf{v}_2 \mathbf{v}_2 d\mathbf{v}_2 \\ &= \frac{2}{15} \left\{ m_1 \int \varphi_1^{(4)} v_1^4 d\mathbf{v}_1 + m_2 \int \varphi_2^{(4)} v_2^4 d\mathbf{v}_2 \right\} \mathbf{b}, \\ \mathbf{L}' &= - \frac{1}{2} \int \mathbf{r}_1 \frac{\partial \phi_1}{\partial \mathbf{r}_1} N'_{20} d\mathbf{r}_1 - \frac{m_2}{m_1 + m_2} \int \mathbf{s}_1 \frac{\partial \psi}{\partial \mathbf{s}_1} N'_{11} d\mathbf{s}_1 \\ &\quad - \frac{1}{2} \int \mathbf{r}_2 \frac{\partial \phi_2}{\partial \mathbf{r}_2} N'_{02} d\mathbf{r}_2 - \frac{m_1}{m_1 + m_2} \int \mathbf{s}_2 \frac{\partial \psi}{\partial \mathbf{s}_2} N'_{11} d\mathbf{s}_2 \\ &= - \frac{1}{15} \left\{ \int v_{20}^{(4)} \phi_1'(r_1) r_1^3 d\mathbf{r}_1 + \int v_{02}^{(4)} \phi_2'(r_2) r_2^3 d\mathbf{r}_2 + 2 \int v_{11}^{(4)} \psi'(s_1) s_1^3 d\mathbf{s}_1 \right\} \mathbf{b}, \end{aligned} \right\} \quad (87)$$

where the relations (50), (45), (83), (80) and (82)' have been used.

According to the definition of the coefficient of viscosity μ

$$p' = -2\mu b, \quad (88)$$

one has therefore

$$\left. \begin{aligned} \mu &= \mu_p + \mu_k, \\ \mu_p &= \frac{1}{30} \left\{ \int \nu_{20}^{(4)} \phi_1'(r_1) r_1^3 d\mathbf{r}_1 + \int \nu_{02}^{(4)} \phi_2'(r_2) r_2^3 d\mathbf{r}_2 + 2 \int \nu_{11}^{(4)} \psi'(s_1) s_1^3 d\mathbf{s}_1 \right\}, \\ \mu_k &= -\frac{1}{15} \left\{ m_1 \int \varphi_1^{(4)} v_1^4 d\mathbf{v}_1 + m_2 \int \varphi_2^{(4)} v_2^4 d\mathbf{v}_2 \right\}, \end{aligned} \right\} \quad (89)$$

where the first part is the potential viscosity and the second part the kinetic viscosity, the former playing a predominant role in liquid mixtures. As mentioned in § 2 of part I, diffusion is certainly the simplest transport process of all, as it involves only the kinetic transfer of number densities. The present calculation shows that only F'_1 and F'_2 are required for obtaining the coefficients of diffusion. The practical method of solution for (75) discussed in § 6 indicates a possible way of solution to the problem of diffusion in liquid mixtures which has hitherto been regarded as almost impossible.

The author wishes to express his thanks to Professor M. Born for his interest and encouragement and to Dr H. S. Green for many helpful suggestions and discussions and for reading the manuscript.

REFERENCES

- Born, M. & Green, H. S. 1946 *Proc. Roy. Soc. A*, **188**, 10.
 Born, M. & Green, H. S. 1947a *Proc. Roy. Soc. A*, **190**, 455.
 Born, M. & Green, H. S. 1947b *Proc. Roy. Soc. A*, **191**, 168.
 Chapman, S. & Cowling, T. G. 1939 *The mathematical theory of non-uniform gases*. Cambridge University Press.
 Frenkel, J. 1946 *Kinetic theory of liquids*. Oxford University Press.
 Kirkwood, J. G. 1946 *J. Chem. Phys.* **14**, 180.
 Kirkwood, J. G. 1947 *J. Chem. Phys.* **15**, 72.
 Rodriguez, E. A. 1949 *Proc. Roy. Soc. A*, **196**, 73-92.

The behaviour of continuous stanchions

By J. F. BAKER, M. R. HORNE AND J. W. RODERICK

Department of Engineering, University of Cambridge

(Communicated by Sir Geoffrey Taylor, F.R.S.—Received 17 December 1948—

Revised 14 March 1949)

A description is given of tests carried out on small-scale steel stanchions of rectangular and Γ -section subjected to the arrangements of load encountered in building frames. Two main types of loading are distinguished according to whether the stanchion is bent in single or double curvature. A theoretical explanation of the results observed is sought by reference to the simple plastic theory in which it is assumed that sections plane before bending remain plane after bending. The theory of members subjected to combined bending and axial load in the partially plastic range is developed, and is applied to the case of single curvature stanchions. The growth of the plastic zones is traced up to the stage at which complete plasticity occurs at three sections in the stanchion, and satisfactory agreement is obtained between the theoretical and observed collapse load. When the simple plastic theory is applied to double-curvature bending, inaccuracies arise in certain cases due to strain reduction in the plastic zones, and the simple theory is therefore elaborated to take account of the irreversible nature of plastic strains. The improved theory is then applied to appropriate cases of double-curvature bending.

1. INTRODUCTION

Despite the fact that pin-ended struts are rarely used in practical structures, investigators have confined their attention almost entirely to the behaviour of this simple form of compression member. Various attempts have been made to apply the knowledge thus gained to the design of the more usual member which is continuous through a number of stories or panels. In Great Britain it is usual to treat each length between floors or panel points as equivalent to a pin-ended strut of some effective length depending on the rigidity of the members connected to the continuous compression member. This has been shown to be irrational (Baker 1934). A full description of the loading conditions to which a continuous stanchion in a building frame can be subjected, and a rational method of design, have been given in the *Final Report of the Steel Structures Research Committee* (Baker & Holder 1936; Baker & Williams 1936). That investigation was, however, confined to the elastic range; little or nothing is known of the conditions under which collapse occurs.

An account is given here of an experimental investigation on small-scale steel frames with rigidly connected members to study the collapse of stanchions under various conditions of loading. A description is also given of an analysis which enables the behaviour of the stanchion to be followed, not only in the elastic range, but also in the plastic range right up to collapse.

2. TESTS ON STANCHIONS BENT IN SINGLE AND DOUBLE CURVATURE

In general, a stanchion length in a building frame will be subjected to axial load and end-bending moments. In the majority of cases the end-moments will be of the same sense, so bending the stanchion in double curvature, but in some cases the end-

moments can be of opposite sense, so bending the length in single curvature. In the tests carried out these conditions have been simulated in the frames shown in figure 1.

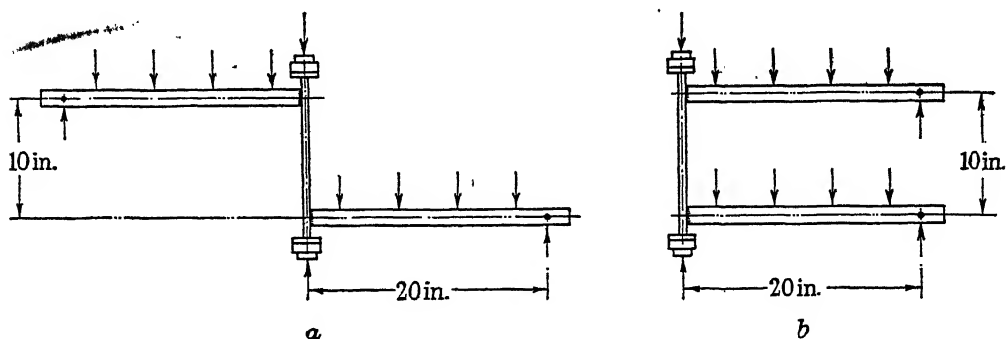


FIGURE 1. Small-scale test frames for stanchions bent in (a) single and (b) double curvature.

When the beams are loaded the stanchion at *a* is bent in single curvature and that at *b* in double curvature. These frames were tested in special loading gear, in which a given load was first applied to each beam. The test was completed by applying to the stanchion an increasing axial load until collapse occurred. To ensure that failure was confined to the stanchion the beams of each test frame were of heavy rectangular section, $1\frac{1}{2}$ in. deep \times $\frac{3}{4}$ in. wide, and were of high tensile steel. The stanchions themselves were of normalized mild steel, the cross-sectional dimensions of each of the four types tested being shown in figure 2.

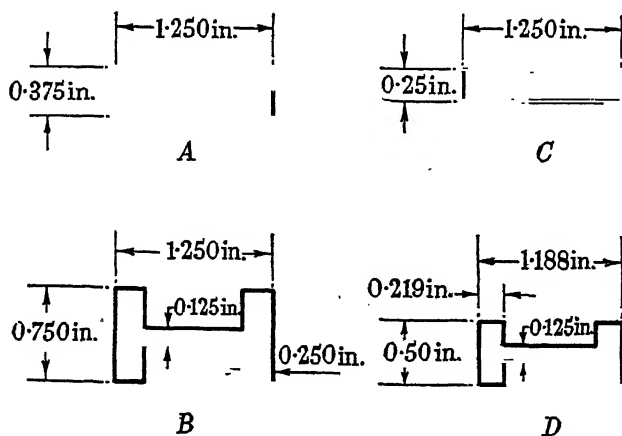


FIGURE 2. Stanchion sections.

The test results for rectangular and Γ -section stanchions were similar in character. Typical load-deflexion curves are shown at *a* and *b*, figure 3, for the heavier stanchions of rectangular section tested in single and in double curvature respectively. At *a* the central deflexion of the stanchion relative to its ends has been plotted against axial load, and for the double-curvature case *b*, the deflexions are those at the quarter points. In both cases the range *OA* represents the deformation produced by beam loading only, and that beyond by the additional axial loading. In figure 4 values of

the axial load to cause collapse (curve 1) and those to develop an extreme fibre stress equal to the lower yield value for the material (curve 2) have been plotted against beam loads for those frames having stanchions of heavier Γ -section. The curves at *a* refer to the single-curvature tests and those at *b* to the double-curvature tests. The horizontal line (e.g. at 8.03 tons in *a*) represents the load which would

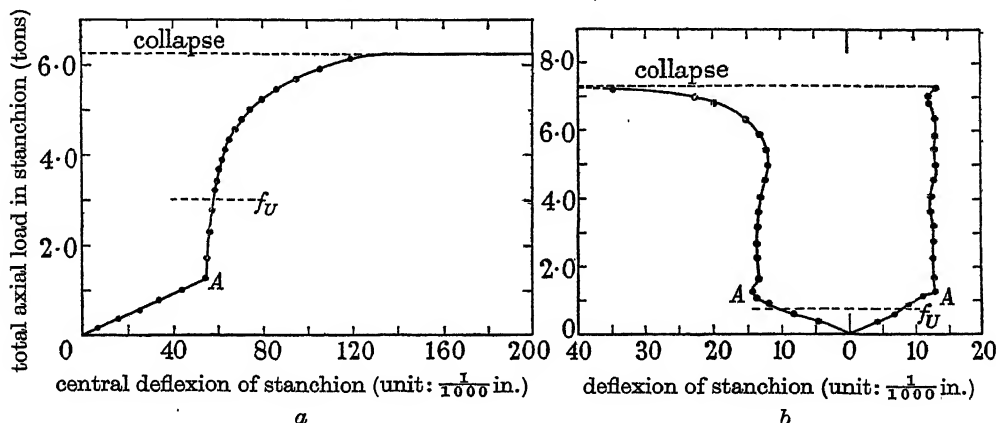


FIGURE 3. Typical load-deflection curves for stanchions of $1\frac{1}{4} \times \frac{3}{8}$ in. rectangular section subjected to (a) single- and (b) double-curvature bending.

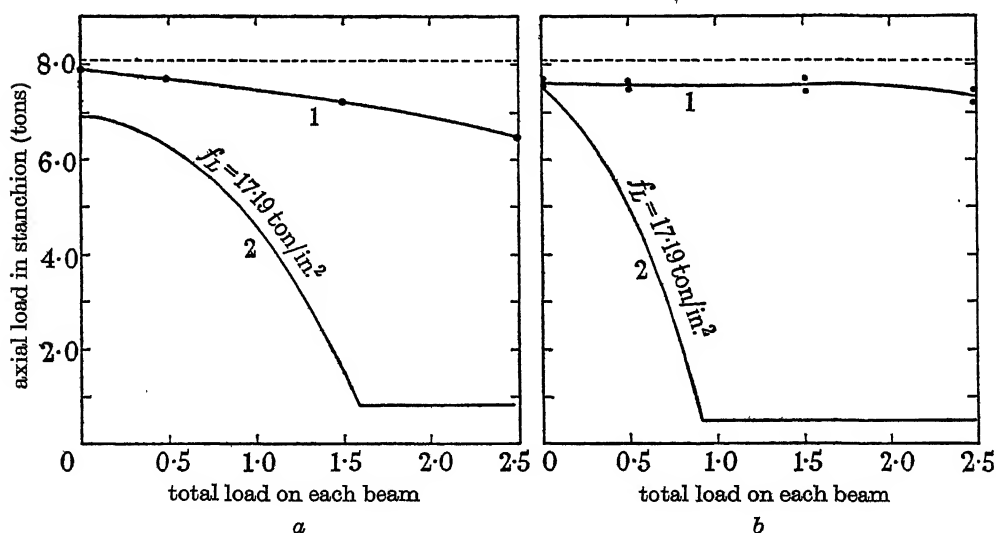


FIGURE 4. Relationship between axial load in stanchion and total load on each beam for stanchions of heavy Γ -section, subjected to (a) single- and (b) double-curvature bending. Curve 1, axial load at collapse. Curve 2, axial load at beginning of yield.

produce failure of the stanchion in pure compression, the corresponding lower yield stress being marked on curve 2. In both single and double curvature, particularly in the latter case, as the frames were subjected to heavier beam loads, the axial load to develop lower yield stress fell away much more rapidly than the collapse load. This is of particular significance in the design of rigidly jointed structures, since curve 2 represents the axial load which is usually regarded as the safe limit for

a continuous stanchion. For instance, from the curves at *b*, for a frame carrying a beam load of 2.5 tons this limiting axial load is 0.44 ton, whereas the stanchion actually withstood a load approximately 17 times greater.

3. THE EQUATIONS OF BENDING FOR A RECTANGULAR STANCHION STRESSED BEYOND THE YIELD POINT

Before the collapse load of a continuous stanchion can be determined analytically it is necessary to know the distribution of stress across a section produced by bending moment and axial load of sufficient magnitude to cause yielding (Roderick 1948). Let

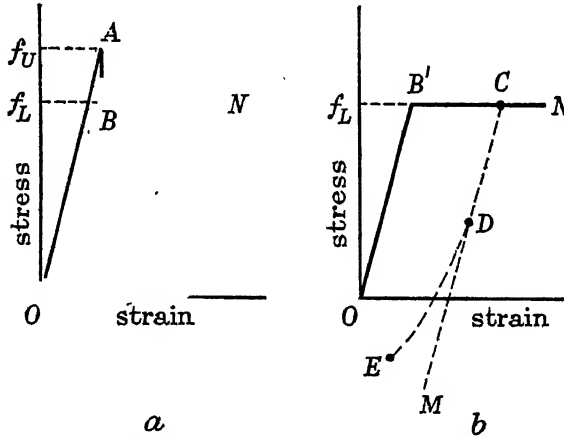


FIGURE 5. Idealized stress-strain relationships of mild steel.

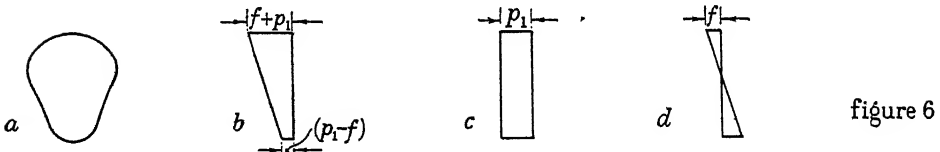


figure 6

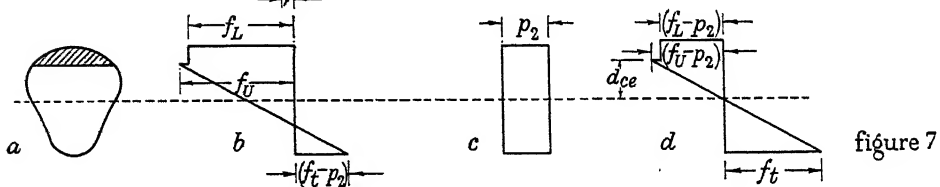


figure 7

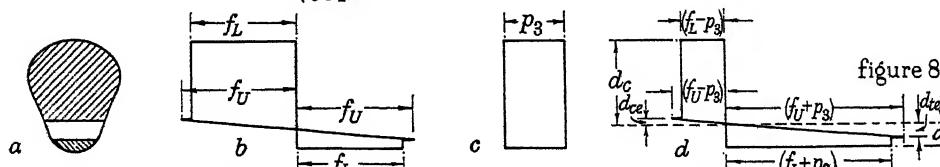


figure 8

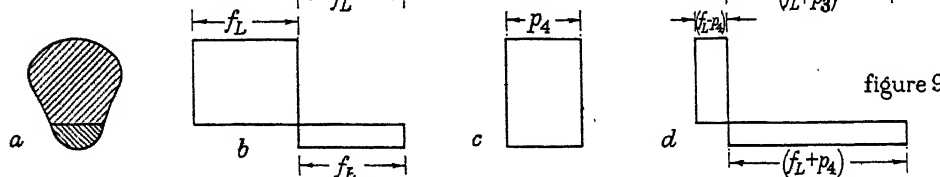


figure 9

FIGURES 6 to 9. Stress distribution in a member subjected to bending moment and axial load for the elastic state (figure 6) and various degrees of plasticity (figures 7 to 9).

the stress-strain relationship for the material be that given in figure 5*a*, where f_U and f_L represent the upper and lower yield stresses respectively. Referring to figure 6, consider the section a of a member subjected to an axial compressive force P and bending moment M , then for elastic conditions,

$$P = A p_1 \quad (1)$$

and

$$M = f Z, \quad (2)$$

where A is the area, and Z the modulus of section. The resultant stress distribution b is therefore made up of a uniform compressive distribution c and one of varying bending stress d having an extreme value f . It is evident from the resultant distribution that yield will occur first at the extreme fibre where the stress is $f + p_1$. It is assumed that, on increasing the straining action beyond the point where yield occurs, plane sections remain plane and that each fibre as it reaches the upper yield stress f_U , drops back to the lower yield value f_L , and thereafter remains constant, the contention being that the fibres are never sufficiently extended to cause strain hardening. Accordingly, as in figure 7, after yield, a resultant stress distribution b is obtained where the material at the one extreme fibre is plastic while at the other, the stress, having decreased, eventually changes sign. If now the axial stress p_2 is deducted from the total distribution, the bending stress distribution d is obtained. The condition represented by the diagrams of figure 7 ceases to apply when the resultant stress $f_i - p_2$ eventually reaches the upper yield value for the material, so giving rise to the distributions shown in figure 8. This is the last stage in the development of the moment of resistance, and in the course of it more and more fibres on both the tension and compression sides reach the upper yield stress and revert to the lower yield value, until finally the whole section is in a state of plasticity (figure 9).

In expressing the stress distributions of figures 7 to 9 analytically, various conditions have to be satisfied. The total axial load P acting at the section is balanced by the uniform stress distribution shown in figures c , and three equations similar to equation (1) are obtained. The stress distributions in figures d therefore represent bending moment only, and equating the total axial force to zero,

$$d_{ce}\{(f_L + p) A_t - (f_L - p) A_c\} = (f_U - p) \{A_\gamma \bar{y}_\gamma - A_\tau \bar{y}_\tau\}, \quad (3)$$

where p = mean axial stress,

A_t, A_τ = area of plastic and elastic regions respectively on the tension side,

A_c, A_γ = area of plastic and elastic regions respectively on the compression side,

\bar{y}_τ = distance of centre of gravity of tensile elastic region from the neutral axis,

and \bar{y}_γ = distance of centre of gravity of compressive elastic region from the neutral axis.

The above equation refers strictly to figure 8*d* only, but corresponding equations for figures 7*d* and 9*d* can be obtained by omitting appropriate terms.

If the bending moment acting on the section is denoted by M , then it follows by taking moments about the neutral axis in figure 8*d* that

$$M = (f_U - p) Z_e + (f_L - p) Z_{pc} + (f_L + p) Z_{pt}, \quad (4)$$

where $Z_e = I_e/d_{ce}$, I_e being the second moment of area of the elastic region,

Z_{pt} = the first moment of area of the plastic region on the tension side,

and Z_{pc} = the first moment of area of the plastic region on the compression side.

Again, corresponding equations for figures 7*d* and 9*d* may be obtained by the omission of the appropriate terms.

All the equations given above refer to a member of any section which has an axis of symmetry in the plane of bending. Attention is hereafter confined to members of rectangular cross-section. The width will be denoted by b and the depth by $2d$. Equation (1) becomes

$$P = 2bdp, \quad (5)$$

and is applicable to all stress distributions. For the stress distributions in figure 6, the expression for the bending moment as given by equation (2) becomes

$$M = \frac{2}{3}bd^2f. \quad (6)$$

The stress distribution in figure 7*d* is adequately defined by the values of p , f_t and d_{ce} . Expressed in terms of these variables, equation (3) leads to

$$d_{ce} = \frac{2d}{\{av^2 + v + (1-a)\}}, \quad (7)$$

where

$$v = \frac{f_t}{(f_U - p)}$$

and

$$a = \frac{(f_U - p)}{2(f_L - p)}.$$

It is found convenient to introduce a parameter ϕ such that

$$\sec \phi = \frac{2av + 1}{2a - 1}.$$

It then follows from equation (7) that

$$d_{ce} = \frac{8ad}{(2a-1)^2} \cot^2 \phi. \quad (8)$$

By substituting in equation (4) it is ultimately found that

$$M = \frac{2}{3}bd^2 \frac{(f_L - p)}{(2a - 1)} \{3(2a - 1)(\sec^2 \phi - 2)\sec^2 \phi - 4(\sec^2 \phi - 3)\sec \phi + (6a + 5)\} \cot^4 \phi. \quad (9)$$

The stress distribution in figure 8*d* is adequately defined by the values of p , d_e and d_{ce} . If equation (3) is expressed in terms of these variables, it is found that

$$d_e = \frac{(f_L + p)}{f_L} d - \frac{p}{(f_U - p)} d_{ce}. \quad (10)$$

The bending moment given by equation (4) may be expressed in terms of p , d_e and d_{ce} , and after substitution for d_e from equation (10) it is found that

$$M = \frac{(f_L^2 - p^2)}{f_L} bd^2 - \frac{f_U(3f_L - 2f_U)}{3(f_U - p)^2} bd_{ce}^2. \quad (11)$$

The stress distribution in figure 9*d* corresponds to the condition of full plasticity, and is the limiting case of figure 8*d* when $d_{ce} = 0$. The value of the bending moment at full plasticity is therefore given by substituting $d_{ce} = 0$ in equation (11).

It will be seen from figures 6 to 9 that when a member is subjected to bending about one principal axis combined with axial load, plastic zones can be formed as shown in figure 10. In the length AB , to be referred to as the elastic length, the stress

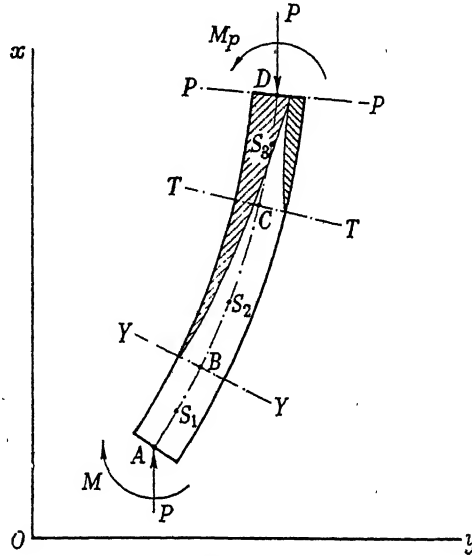


FIGURE 10

distributions are of the type shown in figure 6*b*. At section YY yield stress is just reached in the extreme compression fibres, and hence from figure 6, $f = (f_U - p)$. The bending moment then has the particular value M_y as given by substituting this value of f in equation (6). Hence

$$M_y = \frac{2}{3}(f_U - p)bd^2. \quad (12)$$

In the length BC , to be referred to as the primary plastic length, the stress distributions are of the type shown in figure 7*b*. At section TT , yield stress is just reached in the extreme tension fibres, and hence from figure 7, $f_t = (f_U + p)$. The bending moment has the particular value M_T as given by substituting this value of f_t in the expressions leading to equation (9). It is thus found that

$$M_T = \frac{2}{3} \frac{(f_L - p)}{f_L^2} \{f_U(f_L - p) + 3f_L p\} bd^2. \quad (13)$$

Finally in the length CD , to be referred to as the secondary plastic length, the stress distributions are of the type shown in figure 8*b*, until complete plasticity is reached at section PP . The value M_p of the bending moment at this last section is readily obtained by substituting $d_{ce} = 0$ in equation (11) and hence

$$M_p = \frac{(f_L^2 - p^2)}{f_L} bd^2. \quad (14)$$

Having obtained expressions for the moment of resistance at any section, it is possible to derive slope-deflexion equations, separate sets of equations being necessary for each of the elastic and plastic lengths. It is convenient to take rectangular axes such that the Ox axis is coincident with the line of action of the resultant forces acting on the member. Thus Ox in figure 10 is coincident with the line of action of those equal and opposite forces which are the resultants of P , M at A and P , M_p at D . The position of the Oy axis is arbitrary. It will be assumed that the shear forces in the member are everywhere small, so that the angle between Ox and the longitudinal axis of the member is small. Hence if the radius of curvature R of the centre line is large compared with the length of the member,

$$\frac{d^2y}{dx^2} = -\frac{1}{R}, \quad (15)$$

where (x, y) are the co-ordinates of a point on the central axis. If M is the bending moment at any section, then

$$M = Py. \quad (16)$$

Considering first the elastic length, it follows from the stress distribution, figure 6*d*, that

$$\frac{1}{R} = \frac{f}{Ed}, \quad (17)$$

where E is the modulus of elasticity. It therefore follows from equations (6), (15), (16) and (17) that

$$EI \frac{d^2y}{dx^2} + Py = 0, \quad (18)$$

where I is the moment of inertia of the cross-section of the beam about the axis of bending and has the value $\frac{2}{3}bd^3$. Let x and dy/dx have the known values x_1 and θ_1 at any section S_1 in the elastic length, and let the bending moment there be M_1 . If equation (18) is integrated and the constants of integration evaluated by reference to point S_1 , it is readily shown that at any section,

$$\frac{dy}{dx} = \theta_1 \cos \alpha(x-x_1) - \alpha \frac{M_1}{P} \sin \alpha(x-x_1), \quad (19)$$

$$y = \frac{1}{\alpha} \theta_1 \sin \alpha(x-x_1) + \frac{M_1}{P} \cos \alpha(x-x_1), \quad (20)$$

where

$$\alpha^2 = \frac{P}{EI}.$$

In the case of the primary plastic length, it follows from the geometry of the stress-strain diagram in figure 7*d* that

$$\frac{1}{R} = \frac{(f_U - p)}{Ed_{ce}}. \quad (21)$$

The combination of equations (9), (15), (16) and (21) leads to a differential equation which may be integrated once. Let x and dy/dx have the known values of x_2 and θ_2

at any section S_2 in the primary plastic length, and let the parameter ϕ have the value ϕ_2 . Then the constant of integration is readily evaluated, and it is found that

$$\frac{dy}{dx} = \sqrt{[\theta_2^2 - \beta(\epsilon - \epsilon_2)]}, \quad (22)$$

where

$$\beta = \frac{2(f_U - f_L)(f_L - p)}{3Ep},$$

$$\epsilon = 4 \cot \phi \cot \frac{1}{2}\phi + \sec \phi,$$

$$\epsilon_2 = 4 \cot \phi_2 \cot \frac{1}{2}\phi_2 + \sec \phi_2.$$

Equation (22) cannot be integrated directly, and it is necessary to resort to a graphical solution. If, however, no account is taken of the upper yield stress (figure 5b), the various expressions will be simplified by putting $f_U = f_L$. In these circumstances, equation (9) reduces to

$$M = \frac{3v-1}{v+1} M_y, \quad (23)$$

where M_y has the value given in equation (12). The differential equation obtained from equations (15), (16), (21) and (23) can be integrated twice without difficulty. If the value of V at S_2 is denoted by V_2 , then

$$\frac{dy}{dx} = \sqrt{[\theta_2^2 - \mu(v - v_2)]}, \quad (24)$$

$$x - x_2 = \frac{E\mu d}{c^2(f_L - p)} \left\{ 2 \left[\frac{\theta}{v+1} - \frac{\theta_2}{v_2+1} \right] + \frac{\mu}{c} \log \frac{(c+\theta)(c-\theta_2)}{(c-\theta)(c+\theta_2)} \right\}, \quad (25)$$

where

$$\mu = \frac{2(f_L - p)^2}{3Ep}$$

and

$$c = \sqrt{[\theta_2^2 + \mu(1 + v_2)]}.$$

Considering the secondary plastic length, equation (21) still applies, and leads, in combination with equations (11), (15) and (16), to a differential equation which may be integrated directly. If the values of x and dy/dx are x_3 and θ_3 at some point S_3 in the secondary plastic length where the value of d_{ce} is given by d_3 , then it is found that

$$\frac{dy}{dx} = \sqrt{\left[\frac{4(f_U - p)}{Em} (d_{ce} - d_3) + \theta_3^2 \right]}, \quad (26)$$

$$x - x_3 = \frac{2}{3s} \sqrt{\left[\frac{E}{ms(f_U - p)} \right]} \{ (2 - sd_3) \sqrt{(sd_3 + 1)} - (2 - sd_{ce}) \sqrt{(sd_{ce} + 1)} \}, \quad (27)$$

where

$$m = \frac{6(f_U - p)^2 p d}{f_U^2 (3f_L - 2f_U)},$$

and

$$s = \frac{4(f_U - p)}{\{Em\theta_3^2 - 4(f_U - p)d_3\}}.$$

4. THE THEORETICAL BEHAVIOUR OF A STANCHION BENT IN SINGLE CURVATURE

To show the use of these equations a single-curvature frame will be examined (figure 1a) having a stanchion of the section denoted by *A* in figure 2 and made from a steel with an upper-yield stress of 22.88 tons/sq.in. and a lower-yield value of 20.32 tons/sq.in. Each beam carried a total load of 1.99 tons, and collapse occurred when the axial load in the stanchion was 6.67 tons. From ordinary elastic theory it was found that yield must have occurred in the extreme fibre of the section at the

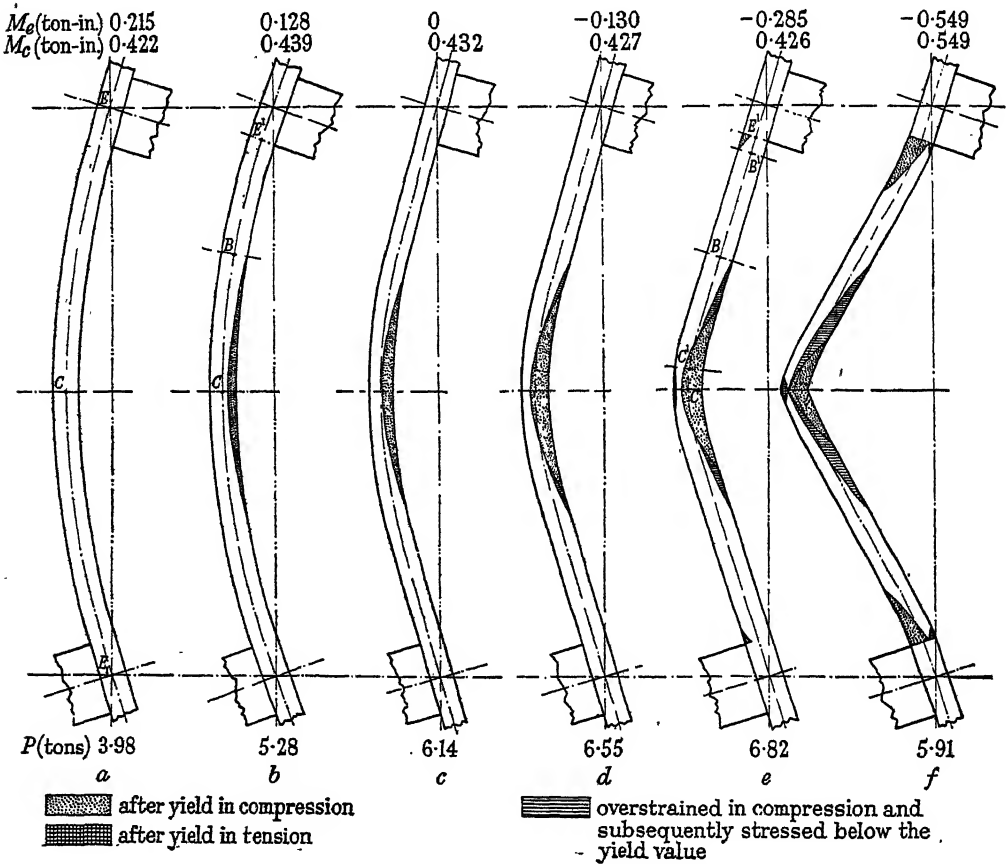


FIGURE 11. The deflected form of a stanchion bent in single curvature at various stages of loading, showing the extent of the plastic zones.

centre of length of the stanchion when the axial load was 3.98 tons. The calculated deflected form of the stanchion at this stage is shown in figure 11a, the deflexions being drawn to a scale 12.5 times the linear scale. The corresponding deflexion at the centre *C* (figure 11) is denoted by *N* in figure 12, where this quantity is plotted against the axial load in the stanchion.

After yielding has taken place further calculation is only possible with the aid of equations (19) to (27). A process of trial and error has to be introduced. For a given axial load a value for the bending moment transmitted from beam to stanchion is

assumed. This makes it possible to calculate the change of slope at the end of the beam, and with this as a starting point the lengths of the elastic portion $E'B$ (figure 11*b*) and of the primary plastic length BC of the stanchion can be determined. When the value of the moment at the connexion has been chosen correctly the sum of the lengths $E'B$ and BC will be equal to the known half-length (4.375 in.) of the stanchion. The diagrams *c* and *d* in figure 11 were obtained in a similar manner.

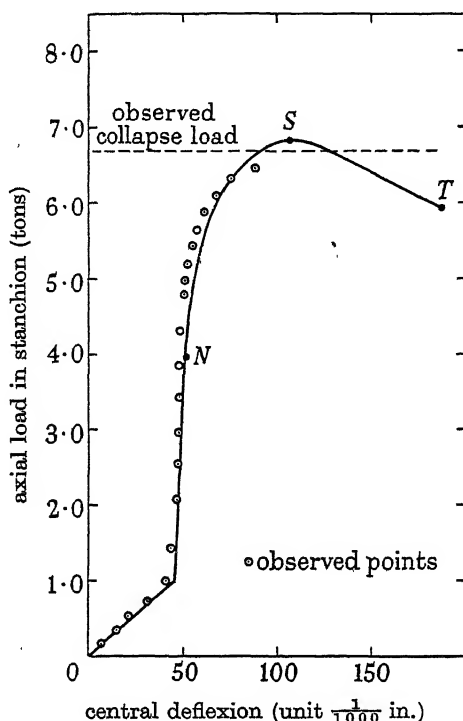


FIGURE 12. Comparison of observed and theoretical relationships between total axial load and central deflection for a stanchion bent in single curvature.

It will be seen from the information given above each diagram (figure 11) that the end-moment (M_E) decreased steadily with increase in axial load. In the test this had the effect of preventing the growth of the plastic zone along the length of the stanchion, despite the fact that its depth was increasing continuously at the centre. The effect was so marked for an axial load of 6.55 tons (diagram *d*) that some part of the yielded material at the extremities of what had been the plastic zone must have been subjected to stresses lower than the yield value. These regions are indicated according to the key in figure 11.

From these calculations it was found that yielding occurred on the tension face of the stanchion at the centre of its length at an axial load of 6.75 tons. This was followed shortly afterwards, at an axial load of 6.78 tons, by yielding at the ends of the stanchion on the compression face. In diagram *e* of figure 11 it will be seen that, at an axial load of 6.82 tons, there is a primary plastic length $E'B'$, an elastic length $B'B$, another primary plastic length BC' , and finally a secondary plastic length $C'C$.

Under these conditions the calculated central deflexion of the stanchion has the value indicated by S in figure 12.

Any attempt to consider a more advanced condition of plasticity fails to give a solution, unless in making the calculation the axial load P is assumed to have a value less than 6.82 tons. Proceeding in this manner it is possible to show that the load deflexion diagram beyond the point S is given by ST in figure 12. In this range the depths of the plastic zones at C (figure 11*e*) increase until the tensile and compressive zones meet and the section becomes completely plastic. Thereafter it is necessary to assume that what has been called a 'plastic hinge' occurs at this section so that further straining results in rotation. At the same time the degree of plasticity increases at the ends until yield also occurs on the tension face, the corresponding value of the axial load being 6.37 tons, and complete plasticity occurs at the ends under an axial load of 5.91 tons (figure 11*f*).

In comparing these theoretical results with the observed values (figure 12) it must be remembered that, since dead loading was used in the test it was not possible to reduce the axial load so that the range ST could be followed. Collapse therefore took place at an axial load (observed to be 6.67 tons) corresponding to the point S . The observed deflexion points are seen to be in fairly good agreement with the theoretical curve up to the collapse load.

A subsequent calculation assuming $f_U = f_L$ showed that this approximation introduces little error in deflexion or collapse load.

This simple theory, however, is not satisfactory for double-curvature bending when plastic zones are developed in the course of beam loading. When axial load is applied the resultant decrease in end-moment reduces the strain in the plastic zones, the corresponding stress-strain relationship being represented approximately by CM in figure 5*b*, while neglect of the effects of overstrain is equivalent to assuming that plastic strain is reversible and represented by $CB'O$. In order to develop a more exact analysis the plastic stress distributions at every section have to be modified at each stage to take account of their strain history. The principles underlying this process are demonstrated in § 5 below.

5. THE EQUATIONS OF BENDING FOR A RECTANGULAR SECTION WHEN STRESS REVERSAL OCCURS IN THE PLASTIC RANGE

The process of following the changes in stress of all the fibres in the cross-section of a stanchion stressed beyond the yield point is complicated, since a large variety of cases arise. It is therefore proposed to consider one example only in demonstrating the method which has been used.

Consider a member of rectangular section (figure 13*a*) of width b and depth $2d$ subjected to the forces shown in figure 10. If it is assumed that $f_U = f_L$ the stress distribution in the secondary plastic length will be that represented by $GH CDEF$ in figure 13*b*. As long as the straining actions are increasing after the yield point has been reached in any fibre, all subsequent changes of strain are of the same sign and no decrease of stress occurs. The relationship between M , the bending moment, P ,

the axial load, and R , the radius of curvature in the plane of bending, may therefore be obtained from equations (5), (11) and (21) above. The line CD (figure 13*b*) represents the stress distribution in the elastic part of the cross-section, and hence the same line cd produced to a and b as in figure 13*c* represents, to some scale, the longitudinal strain over the whole section. Let the extreme fibre strains in compression and tension, represented by points a and b , be denoted by ϵ_c and ϵ_t respectively. The line CD in the stress diagram b may also be extended to A and B , giving intercepts of length g_c and g_t respectively, then

$$g_c = E\epsilon_c, \quad g_t = E\epsilon_t \quad \text{and} \quad f_L = E\epsilon_f.$$

Now consider what happens when the applied bending moment changes to M_1 and the axial load to P_1 in such a way that the longitudinal strain in all fibres of the cross-section is represented by the line a_1b_1 in figure 13*c*. Let a_1b_1 intersect ab at a point O between a and c . Let the extreme fibre strains in compression and tension,

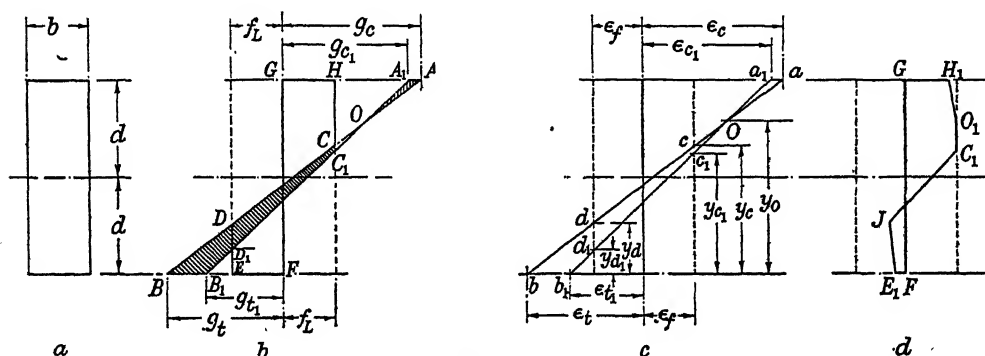


FIGURE 13. Stress-strain diagram for a member subjected to stress reduction in the partially plastic range.

as indicated by points a_1 and b_1 , be denoted by ϵ_{c_1} and ϵ_{t_1} respectively, where $\epsilon_{c_1} < \epsilon_c$ and $\epsilon_{t_1} < \epsilon_t$. The line A_1B_1 may also be drawn in figure 13*b*, the intercepts g_{c_1} and g_{t_1} being given by $g_{c_1} = E\epsilon_{c_1}$ and $g_{t_1} = E\epsilon_{t_1}$. The material between A and O is now subjected to a positive tensile change of strain, and that between O and B to a negative compressive change of strain. Since the material between A and O is initially stressed beyond the yield point in compression, this positive strain produces a decrease in stress which, since it takes place elastically, is represented by AOA_1 in figure 13*b*. Between O and C the strain consists of plastic deformations and there is no change of stress. Between C and C_1 material initially elastic is stressed beyond the yield point in compression, while between C_1 and D the material remains wholly elastic. Between D and B material initially stressed beyond the yield point in tension is subjected to a negative strain, and the stress is therefore decreased elastically. The change of stress in all fibres between C and B is given by CC_1B_1B in figure 13*b*, and the final stress distribution takes the form $GH_1O_1C_1JE_1F$ in figure 13*d*.

Owing to the change of bending moment and axial load, the radius of curvature in the plane of bending changes from R to R_1 . In order to determine the relationship

between M_1 , P_1 and R_1 it is necessary to consider the change of stress distribution in greater detail. Let

$$\rho = \frac{1}{R} = \frac{\epsilon_c + \epsilon_t}{2d}, \quad \rho_1 = \frac{1}{R_1} = \frac{\epsilon_{c1} + \epsilon_{t1}}{2d},$$

$$\epsilon = \frac{\epsilon_c - \epsilon_t}{2}, \quad \epsilon_1 = \frac{\epsilon_{c1} - \epsilon_{t1}}{2},$$

and

$$\sigma = (\epsilon_f - \epsilon) R, \quad \sigma_1 = (\epsilon_f - \epsilon_1) R_1.$$

Also, it is found from the geometry of figure 13c that

$$\left. \begin{aligned} y_c &= d + \sigma, \\ y_{c1} &= d + \sigma_1, \\ y_0 &= d + \frac{\sigma_1 \rho_1 - \sigma \rho}{\rho_1 - \rho} \end{aligned} \right\}. \quad (28)$$

By considering the change in stress distribution given by the areas OAA_1 and BCC_1B_1 in figure 13b, it follows that

$$P_1 - P = b\{BCC_1B_1 - OAA_1\}, \quad (29)$$

$$\text{and } M - M_1 = \frac{1}{2}b\{y_c(g_t + f_L)(d - \frac{1}{3}y_c) - y_{c1}(g_{t1} + f_L)(d - \frac{1}{3}y_{c1}) \\ + (2d - y_0)(g_c - g_{c1})(d - \frac{1}{3}\overline{2d - y_0})\}. \quad (30)$$

Using the values of y_c , y_{c1} and y_0 given by equations (28) it is found that

$$4d\beta = \frac{\rho_1\rho(\sigma_1 - \sigma)^2}{(\rho_1 - \rho)} - 4d(\sigma_1\rho_1 - \sigma\rho), \quad (31)$$

$$4d^2\gamma = \frac{\rho_1\rho(\sigma_1 - \sigma)^2}{(\rho_1 - \rho)} \left\{ \frac{(\sigma_1\rho_1 - \sigma\rho)}{(\rho_1 - \rho)} + \sigma_1 + \sigma \right\} + 4d^3(\rho_1 - \rho), \quad (32)$$

where

$$\beta = \frac{P_1 - P}{2bdE}, \quad \gamma = \frac{3(M_1 - M)}{2bd^2E}.$$

If M , M_1 , P , P_1 , ρ and σ are known, these equations determine the values of ρ_1 and σ_1 . Unfortunately, they can only be solved by trial and error, but if $(M_1 - M)$ and $(P_1 - P)$ are small compared with M and P respectively, a direct solution is possible. Let $M_1 = M + \delta M$ and $P_1 = P + \delta P$. Then equations (31) and (32) become

$$\frac{2}{bE} \frac{\delta P}{\delta \rho} = \rho^2 \left(\frac{\delta \sigma}{\delta \rho} \right)^2 - 4d\rho \left(\frac{\delta \sigma}{\delta \rho} \right) - 4d\sigma, \quad (33)$$

$$\frac{6}{bE} \frac{\delta M}{\delta \rho} = \rho^3 \left(\frac{\delta \sigma}{\delta \rho} \right)^3 + 3\sigma\rho^2 \left(\frac{\delta \sigma}{\delta \rho} \right)^2 + 4d^3. \quad (34)$$

By eliminating $\delta\rho$ from the left-hand side of equations (33) and (34), it is found that

$$\rho^3 \left(\frac{\delta \sigma}{\delta \rho} \right)^3 + 3\rho^2 \left(\sigma - \frac{\delta M}{\delta P} \right) \left(\frac{\delta \sigma}{\delta \rho} \right)^2 + 12d\rho \left(\frac{\delta M}{\delta P} \right) \left(\frac{\delta \sigma}{\delta \rho} \right) + 4d \left(d^2 + 3\sigma \frac{\delta M}{\delta P} \right) = 0. \quad (35)$$

The value of $\delta\sigma/\delta\rho$ can be obtained from equation (35), and substituted in equation (33) to give the value of $\delta\rho$. Hence the change of curvature can be calculated directly

from the changes in bending moment δM and axial load δP . These equations only apply as long as O , the point of intersection of ab and a_1b_1 (figure 13c), lies between a and c , i.e.

$$y_c < y_0 < 2d.$$

Hence

$$\sigma < \sigma_1 \quad \text{and} \quad \rho(d - \sigma) < \rho_1(d - \sigma_1).$$

It is found that altogether it is necessary to consider ten sets of equations covering all the possible cases of change of stress. Thus the point of intersection O of the strain lines ab and a_1b_1 in figure 13c may fall on ba produced, within ac , within cd , within db or on ab produced, giving five possibilities. For each of these positions of intersection, the inclination of a_1b_1 to the line of zero strain may be greater or smaller than that of ab , resulting in ten cases, each of which gives a different type of stress distribution. In the example quoted above, it has been assumed that the initial stress distribution is one in which overstrain has not already occurred, but since only changes of stress are being considered, it is usually—though not always—possible to apply the same equations to more complicated initial stress distributions.

6. THE THEORETICAL BEHAVIOUR OF A STANCHION BENT IN DOUBLE CURVATURE

Having obtained a relationship between the changes in curvature, axial load and bending moment for any initial stress distribution, it is possible to carry out a step-by-step process of integration along a stanchion, after assuming some definite increase in axial load. This process has been applied to a stanchion bent in double curvature, and the results are illustrated in figure 14.

Diagram *a* of figure 14 corresponds to the beam load at which the lower yield stress is first reached in the stanchion. Up to the full beam loading, figure 14*b*, all strain increments in any fibre are of the same sign, and consequently no overstrain effects are introduced. However, as soon as direct axial load is applied, there is a decrease in stress in practically all the tension fibres, and the greater part of the material in the tensile plastic zone in diagram *b* immediately begins to show overstrain effects. The state of the stanchion when the total axial load is 3.00 tons is represented by diagram *c*, the area *abc* being that initially strained beyond the yield point in tension and subsequently stressed below the yield value.

According to the analysis, as the total axial load increases, the compressive plastic zone extends towards the centre of length of the stanchion. At the same time the terminal bending moments decrease, so causing a decrease in stress in some of the extreme compressive fibres near the ends. This begins to occur first when the total axial load is somewhat under 6.42 tons, and consequently there is a compressive overstrain zone at this load denoted by *def* in diagram *d*. The maximum bending moment no longer occurs at the ends of the stanchion but at sections *K* and *K'* at a distance of 0.62 in. from *E'* and *E'_1* (diagram *a*) respectively. The remaining diagrams *e* and *f* correspond to axial loads of 7.50 and 7.90 tons respectively. It will be seen that as the axial load increases the compressive plastic zones become deeper and continue to approach nearer to the centre of length of the stanchion. The compres-

sive overstrain zones also increase in length. The end bending moments decrease to zero and then increase with reversed sign, being 0.075 ton-in. when the axial load is 7.50 tons and -0.132 ton-in. at 7.90 tons.

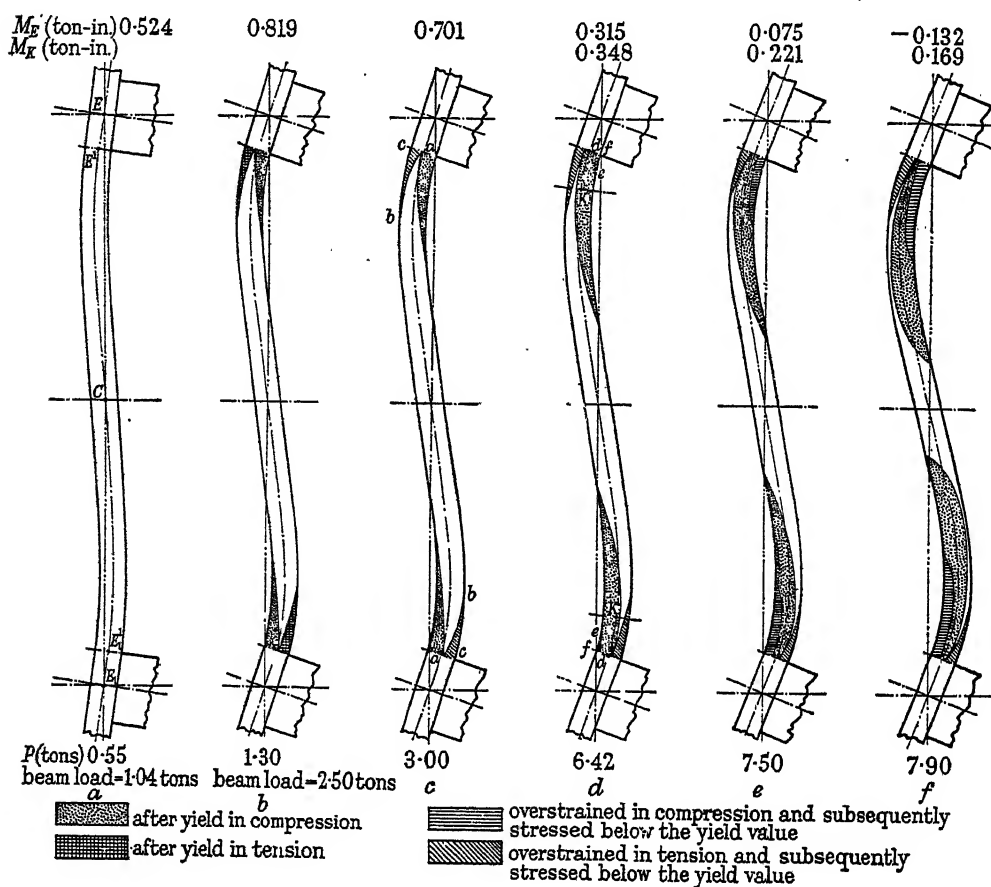


FIGURE 14. The deflected form of a stanchion bent in double curvature at various stages of loading, showing the extent of the plastic zones.

The relationship between axial load and the rotation of the central section is shown by curve 1, figure 15. As the load approaches 7.90 tons the central rotation is increasing rapidly, in fact, to such an extent, that this load could fairly be said to represent the theoretical collapse load. When an analysis was carried out neglecting overstrain, i.e. assuming that the stress-strain relationship shown in figure 5b was reversible, curve 2 was obtained, showing a collapse load of 6.43 tons. These theoretical curves may be compared with the observed points, the observed collapse load being 7.33 tons. Thus the theoretical collapse load was 12.3 % too low when overstrain was neglected, and 7.9 % too high when overstrain was allowed for. The agreement obtained when overstrain is allowed for is therefore better than when it is neglected, and the remaining discrepancy is possibly due to the fact that the stress-strain relation has been over-simplified. It is possible that in the 'unloading'

range the relation may be more accurately represented by some curve such as *CDE* (figure 5*b*) rather than by the straight line *CM* (Howard & Smith 1925). However, despite this, the work carried out does show that the effect of overstrain is appreciable, but that to ignore it gives a calculated collapse load on the safe side.

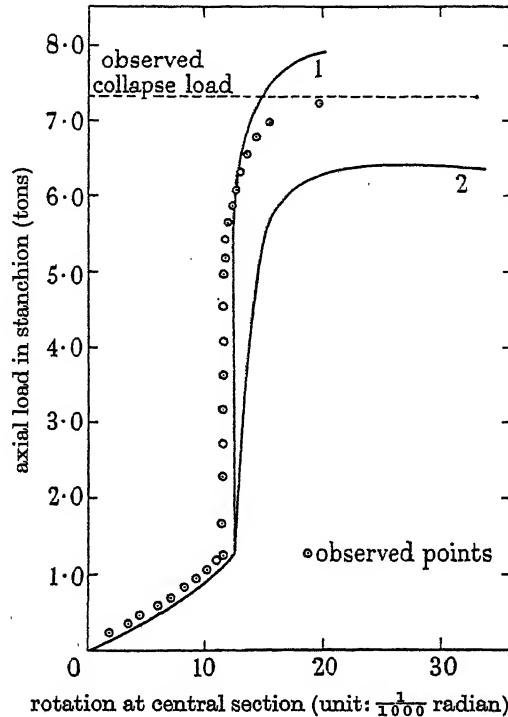


FIGURE 15. Comparison of observed and theoretical relationships between total axial load and rotation of the central section for a stanchion bent in double curvature.

REFERENCES

- Baker, J. F. 1934 *Second Report, Steel Structures Research Committee*, H.M.S.O. 13.
 Baker, J. F. & Holder, P. D. 1936 *Final Report, Steel Structures Research Committee*, H.M.S.O. 436.
 Baker, J. F. and Williams, E. L. 1936 *Final Report, Steel Structures Research Committee*, H.M.S.O. 511.
 Howard, J. V. & Smith, S. L. 1925 *Proc. Roy. Soc. A*, **107**, 113.
 Roderick, J. W. 1948 *Phil. Mag.* **39**, 529.

Studies in polymerization

V. The polymerization of vinyl acetate

By G. DIXON-LEWIS, *Courtaulds Limited, Maidenhead, Berks.*

(Communicated by A. H. Wilson, F.R.S.—Received 21 December 1948)

The velocity constants of propagation, transfer and termination in the polymerization of vinyl acetate at -15 and 0° C have been determined by the viscosity method. A comparison of the results with those of other authors has shown that the values for the propagation and termination constants given in table 6 are probably correct to within a factor of 2.

The quantum yield of the initiating reaction has been determined, and was found to be quite small, about 0.04.

INTRODUCTION

In recent years a number of investigations have been carried out on the polymerization of liquid vinyl acetate, but in no case have reliable values for all the velocity constants been determined. Bamford & Dewar (1948*a*), working with styrene, have shown that four types of reaction are important in vinyl polymerizations, which are chain processes—initiation, propagation, transfer and termination. They have been able, by a new viscometric method, to determine all four velocity constants. In the case of vinyl acetate, Burnett & Melville (1947) and Swain & Bartlett (1946) have determined the propagation and termination constants by the rotating sector method of Briers, Chapman & Walters (1926), combined with independent determinations of the rate of chain starting. However, their values for the termination constant differed by a factor of nearly 40, and it seemed that further determinations were desirable. Earlier work by Bagdassarian (1944) has also shown the importance of chain transfer in the polymerization of vinyl acetate, the average molecular weight of the polymer formed by irradiating the liquid at low light intensities being determined primarily by the chain-propagating and chain transfer reactions.

In the present paper, all the velocity constants for the reactions involved in the polymerization of vinyl acetate have been determined by the method of Bamford & Dewar (1948*a*). The quantum yield of the initiating reaction has also been measured.

Purification of vinyl acetate

Commercial vinyl acetate was dried over calcium chloride, and then fractionated through a 20-plate column in an atmosphere of nitrogen. The sample was thoroughly degassed, and all further purification was carried out *in vacuo* ($p < 10^{-4}$ mm.). An all-glass apparatus was used, and at no stage in the purification was the liquid or its vapour allowed to come into contact with any greased joints or taps. The last traces of water were removed from the liquid by standing for 2 hr. over a few pieces of freshly cut lithium, after which the vinyl acetate was distilled and sealed off in a quartz tube fitted with a vacuum breaker.

The final stage of the purification consisted in the removal of the last traces of acetaldehyde, which is present in commercial vinyl acetate, and of the last traces of inhibitor. The removal of the former substance is important, since, not only is it a photosensitizer for the polymerization, but also it is an active chain transfer agent. Its removal was eventually accomplished, together with the removal of inhibitor, by prolonged irradiation with a hot mercury arc, using a Pyrex glass filter, followed by heating the liquid to 60° C for 2 or 3 days. Under these conditions the acetaldehyde was almost entirely responsible for the initiation of the polymerization reaction, and was itself removed in the initiating process. Preliminary experiments in a Pyrex vacuum viscometer of the usual type had shown that vinyl acetate containing traces of acetaldehyde underwent appreciable polymerization at 25° C in the dark. The final product did not do this.

The vinyl acetate was finally distilled *in vacuo* from the quartz tube into the quartz vacuum viscometers in which the rate measurements were carried out. The viscometers were of the type already described by Bamford & Dewar (1949a).

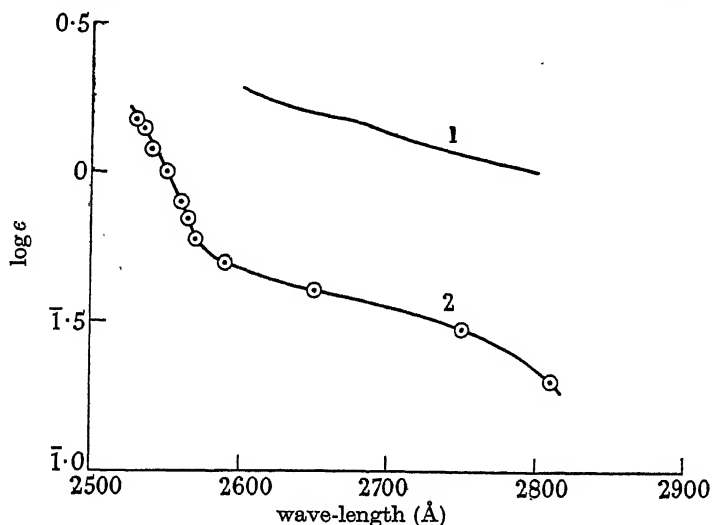


FIGURE 1. The absorption spectrum of vinyl acetate as measured by (1) Burnett & Melville, (2) the author.

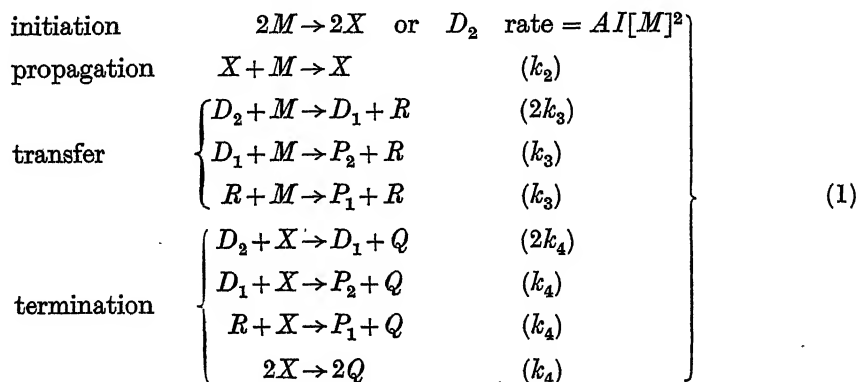
The absorption spectrum of vinyl acetate

The absorption spectrum of a sample of vinyl acetate purified as above, and distilled *in vacuo* into a quartz cell 5 cm. long, was determined in the region of wave-lengths 2900 to 3100 Å by means of a Hilger medium quartz spectrograph and Spekker photometer. The sample had a molar decadic extinction coefficient $\epsilon = 0.014$ at $\lambda = 3000$ Å, whereas at this wave-length acetaldehyde has $\epsilon = 14$ (Schou 1929). The upper limit for the concentration of this substance in the sample is therefore 10^{-3} moles/l.

The absorption spectrum of vinyl acetate in hexane was also determined in a 1 cm. cell using the above instruments. The vinyl acetate had been dried, and then distilled and fractionated in an atmosphere of nitrogen. The spectrum is shown in figure 1, curve 2. It will be noted that the light absorption by the monomer is considerably less than that obtained by Burnett & Melville (1947, figure 1, curve 1).

Summary of the method

The method used to determine the velocity constants in the polymerization is that described by Bamford & Dewar (1948*a*), in which viscosity measurements are used to follow the course of the polymerization. It has already been mentioned that pure vinyl acetate does not polymerize at 25° C in the dark, i.e. $k_1 = 0$, and this fact has made it necessary to modify the method very slightly. The rate of chain starting in the photochemical experiments is simply proportional to the light intensity. The complete reaction scheme is therefore as follows:



where X represents any active centre and Q any dead centre, D_2 is the initial polymer growing at both ends, D_1 is a polymer growing at one end only, R is the growing transfer polymer, and P_1 and P_2 are dead transfer and initial polymers respectively.

In order to interpret the results of the viscosity measurements, it is also necessary to know the relation between viscosity, concentration and mean degree of polymerization. The relations are now known to be as follows:

(a) For homogeneous polymer (Houwink 1940)

$$[\eta] = K(M_0 P)^\alpha, \quad (2)$$

where $[\eta]$ is the intrinsic viscosity, and is equal to the ideal specific viscosity η divided by the concentration of polymer in base moles/l., M_0 is the molecular weight of the monomer, P is the degree of polymerization and K and α are constants.

From (2), the following relations can be derived (Bamford & Dewar 1948*b*):

(b) For heterogeneous polymer produced by the growth of a single radical centre, e.g. for the transfer polymer above,

$$\eta = K'(M_0 \bar{P})^\alpha c, \quad (3)$$

where \bar{P} is the number average degree of polymerization, and

$$K' = K\Gamma(2 + \alpha).$$

(c) For a polymer prepared by the growth of a double radical,

$$\eta = K''(M_0 \bar{P})^\alpha c, \quad (4)$$

where

$$K'' = \frac{K\Gamma(3 + \alpha)}{2^{1+\alpha}} = K' \frac{2 + \alpha}{2^{1+\alpha}} = K' w.$$

Equations (3) and (4) have been used in the present work. For polyvinyl acetate, the results of Wagner (1947) lead to the following values, when c is expressed in the above units, and the solvent is acetone at 25° C:

$$K' = 2.28 \times 10^{-3}, \quad \alpha = 0.68.$$

It should be noted that η in the above equations represents the ideal specific viscosity (i.e. $c[\eta]$, where c is the concentration of polymer). The viscosity values given later are all ideal values. They have been determined from the observed specific viscosities at concentration c by the application of the equation of Schulz & Sing (1943):

$$\eta_c = \frac{(\eta_{sp})_c}{1 + 0.28(\eta_{sp})_c}, \quad (5)$$

which these authors claim to hold for dilute solutions of all polymers. In the present experiments, no measurements were carried out on solutions having a specific viscosity greater than 1. This corresponds to a conversion of less than 1 %.

If it is assumed that the chains are long and that k_2 , k_3 and k_4 are independent of the length of the growing polymer chain, the following two expressions can now be derived:

(a) The number average degree of polymerization is given by

$$\bar{P} = \frac{2k_2}{2k_3 + (k_4 AI)^{\frac{1}{2}}}.$$

Thus, if the polymer is formed by irradiating the monomer by light of very low intensity,

$$\bar{P} = k_2/k_3. \quad (6)$$

(b) The rate of change of the ideal specific viscosity in the photochemical polymerization is given by

$$\frac{d\eta}{dt} = K' M_0^\alpha [M]^2 \left(\frac{k_2}{k_3}\right)^{1+\alpha} \left(\frac{k_3}{k_4}\right)^{1+\alpha} (AI)^{\frac{1}{2}} \frac{k_3 k_4^{-\frac{1}{2}} + 2^\alpha w (AI)^{\frac{1}{2}}}{\{k_3 k_4^{-\frac{1}{2}} + (AI)^{\frac{1}{2}}\}^{1+\alpha}}. \quad (7)$$

With the exception of $k_3/k_4^{\frac{1}{2}}$ and A , all the terms in equation (7) are known. Rate measurements at a number of light intensities enable the above two quantities to be calculated by trial and error. The value of A depends, of course, on the units used for the light intensities. It should be noted that only relative light intensities need be known at this stage; in a later section the absolute light intensities have been determined in order to calculate the quantum yield for the initiating reaction.

In order to determine the three velocity constants absolutely, one further relation between them is required. This is provided by measurement of the photochemical after-effect, or, more strictly, the difference between the after-effect and the pre-effect. This quantity, which will be denoted as $\delta\eta_I^t$, arises from the fact that when the light intensity is altered, e.g. from I to zero, the rate of polymerization does not immediately fall to its dark value, but approaches this latter value gradually. Similar considerations apply when the intensity is varied in the reverse direction. $\delta\eta_I^t$ is defined as the difference between the observed ideal specific viscosity at time t ,

and that calculated on the assumption that the rate changes instantaneously to its final value on turning on or cutting off the light. Experimentally, it can be determined by means of two irradiations at the same intensity, but for different times, or as will be shown later.

The method of calculating $\delta\eta_I^t$ has already been outlined by Bamford & Dewar (1948a); in the present case ($k_1 = 0$), it can be shown that, when $\alpha = \frac{2}{3}$,

$$\begin{aligned} \delta\eta_I^t = & \frac{1}{2} K' M_0^{\frac{2}{3}} \left(\frac{k_2}{k_3} \right)^{\frac{1}{3}} \gamma^5 (AI)^{\frac{1}{3}} k_4^{-\frac{1}{3}} [M] \left[2\sqrt{3} \gamma^{-2} \tan^{-1} \frac{2(\gamma^3 + [X]/\theta) + \gamma}{\sqrt{3} \gamma} \right. \\ & - \gamma^{-2} \log \frac{\{(\gamma^3 + [X]/\theta)^{\frac{1}{3}} - \gamma\}^2}{(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \gamma(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \gamma^2} + \frac{3(2^{\alpha}w - 1)}{(\gamma^3 + [X]/\theta)^{\frac{1}{3}}} \\ & + \frac{\gamma^3 + 2^{\alpha}w}{\lambda^5} \left\{ \frac{1}{2} \log \frac{(\gamma^3 + [X]/\theta) - \nu^3}{(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \lambda(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \lambda^2} \right. \\ & \left. - \sqrt{3} \tan^{-1} \frac{2(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \lambda}{\sqrt{3} \lambda} \right\} \\ & + \frac{\gamma^3 - 2^{\alpha}w}{\nu^5} \left\{ \frac{1}{2} \log \frac{\{(\gamma^3 + [X]/\theta)^{\frac{1}{3}} - \nu\}^2}{(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \nu(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \nu^2} \right. \\ & \left. - \sqrt{3} \tan^{-1} \frac{2(\gamma^3 + [X]/\theta)^{\frac{1}{3}} + \nu}{\sqrt{3} \nu} \right\} \\ & \left. - \frac{3(2^{\alpha}w - 1) \gamma^6}{\nu^3 \lambda^3} \frac{1}{(\gamma^3 + [X]/\theta)^{\frac{1}{3}}} \right]_{[X]^0}^{[X]^t}, \end{aligned} \quad (8)$$

where $\gamma^3 = \frac{k_3}{(k_4 AI)^{\frac{1}{3}}}$, $\lambda^3 = \frac{k_3}{(k_4 AI)^{\frac{1}{3}}} + 1$,

$$\nu^3 = \frac{k_3}{(k_4 AI)^{\frac{1}{3}}} - 1, \quad \theta = \left(\frac{AI}{k_4} \right)^{\frac{1}{3}} [M], \quad [X]_I^t = \frac{\theta}{1 + k_4 \theta t}.$$

Hence, k_2/k_3 , $k_3/k_4^{\frac{1}{3}}$ and A being known, it is possible to determine k_4 from a measurement of $\delta\eta_I^t$ at a known light intensity.

EXPERIMENTAL

(a) Measurement of \bar{P}

Small amounts of polyvinyl acetate were prepared at 0 and at 25° C by irradiating the pure monomer in a quartz tube at low light intensities where, to a sufficient approximation, equation (6) holds.

The light source used was a hot mercury arc combined with, at 0° C, a hot (100° C) 50 % acetic acid filter (see below), and at 25° C a Pyrex glass filter.

Solutions of the polymers so prepared were made up in acetone, and their specific viscosities determined at 25° C with an Ostwald viscometer. The number average degrees of polymerization could then be determined by means of equation (3), which, with substitution of Wagner's values for K' and α , becomes*

$$\eta = 2.28 \times 10^{-3} (M_0 \bar{P})^{\frac{2}{3}} c.$$

* It is necessary to use rational values of α to allow calculation of $\delta\eta$. Compare Bamford & Dewar (1948a). The value $\frac{2}{3}$ has therefore been used instead of the experimental 0.68.

Similar experiments were also carried out on the 0° C polymer using vinyl acetate monomer as solvent, with almost identical results. The values for \bar{P} are given in table 1.

TABLE 1

temperature of polymerization (° C)	$\bar{P} \times 10^{-4}$ exp.	$\bar{P} \times 10^{-4}$ calc. with vinyl acetate as solvent
0	2.045	2.15
25	1.300	—

The last column of table 1 shows that the same value of K' can be used when the solvent is vinyl acetate.

(b) *Measurements of rates*

The photochemical rates were measured at 0 and -15° C for a number of different light intensities. At 0° C the viscometer was enclosed in a quartz tube through which water from an ice-bath was circulated. For the measurements at -15° C a specially designed thermostat tank was constructed with double quartz windows. The viscometer was also clamped in this tank in such a way that the viscosity of the polymer solution could be observed without difficulty. The light source used in both cases was a hot mercury arc combined with the hot 50 % acetic acid filter. Under these conditions very little light of wave-length less than 2650 Å reached the vinyl acetate, and although the 2650 Å line is absorbed by the vinyl acetate rather more strongly than is absolutely desirable to ensure an approximately uniform rate of chain initiation throughout the liquid, it has been necessary to use light of this wave-length in order to produce sufficiently high light intensities for the application of the method.* This procedure does not appear to have introduced any very serious error, but it is possible that the value determined for the termination constant may be somewhat too large. Any error has been removed as far as possible by carrying out the measurements at 0 and at -15° C, where the light absorption would be expected to be smaller than at temperatures above 0° C. Although it was found to be impossible at -15° C to vary the light intensity sufficiently to produce an absolutely unambiguous intensity-rate curve (figure 3), the results obtained are regarded as sufficiently accurate to allow the determination of the constants to within quite narrow limits.

The occurrence of the photochemical after-effect excludes the possibility of measuring the rate under illumination by means of a single exposure. The technique has therefore been to carry out two irradiations at the same light intensity, for t_1 and t_2 sec. respectively. If the viscosity is then measured at such a time after each irradiation that the after-effect is complete, then the difference between the total viscosity increases in the two cases is given by

$$\delta\eta_2 - \delta\eta_1 = a(t_2 - t_1),$$

* Consideration of equation (7) shows that at low light intensities the photochemical rate is proportional to the square root of the intensity, whereas at high light intensities, and when $\alpha = \frac{2}{3}$, it is proportional to $I^{\frac{1}{2}}$. For the application of the method, it is necessary to study the intermediate region of the intensity-rate curve.

where α is the photochemical rate. Theoretically, in the case where there is no dark reaction, the after-effect should be infinite (equation (8) with $[X]_I^t = 0$), but in practice it was found that the viscosity became sensibly constant in less than 20 min. after the end of an exposure. The viscosities were therefore always determined after this time had elapsed.

The relative incident intensities were measured by means of a photocell-galvanometer system.

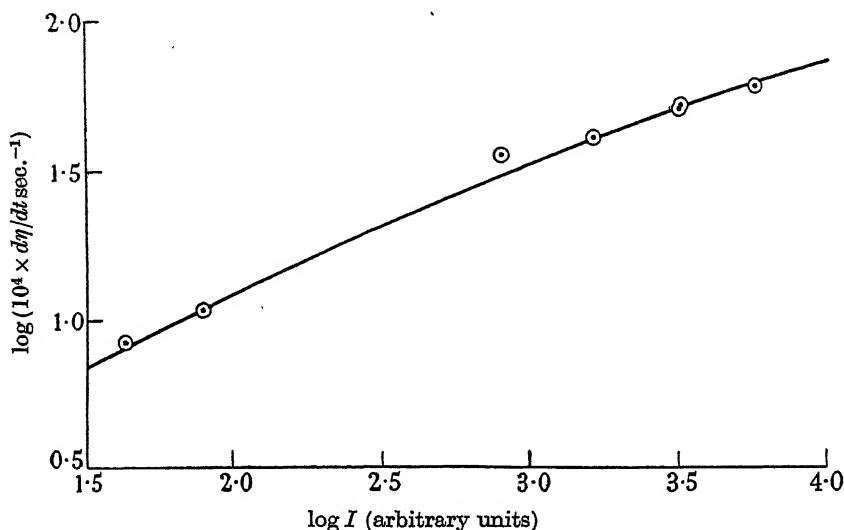


FIGURE 2. The intensity-rate curve for the polymerization of vinyl acetate at 0° C (experimental points and calculated curve).

The results were fitted by trial and error to equation (7), and values of $k_3/k_4^{\frac{1}{2}}$ and A determined, the values of k_2/k_3 being already known (table 1). The values used for K' and α were 2.28×10^{-3} and $\frac{2}{3}$ respectively. The experimentally determined photochemical rates at 0 and -15° C are given in table 2, and the values of the constants giving the best fit in table 3. The experimental points are shown together with the intensity-rate curves calculated from these constants in figures 2 and 3; the agreement is good. The value of k_2/k_3 at -15° C was obtained by extrapolating the curve of $\log \bar{P}$ against $1/T^\circ$ K.

TABLE 2. THE PHOTOCHEMICAL RATES OF POLYMERIZATION OF VINYL ACETATE

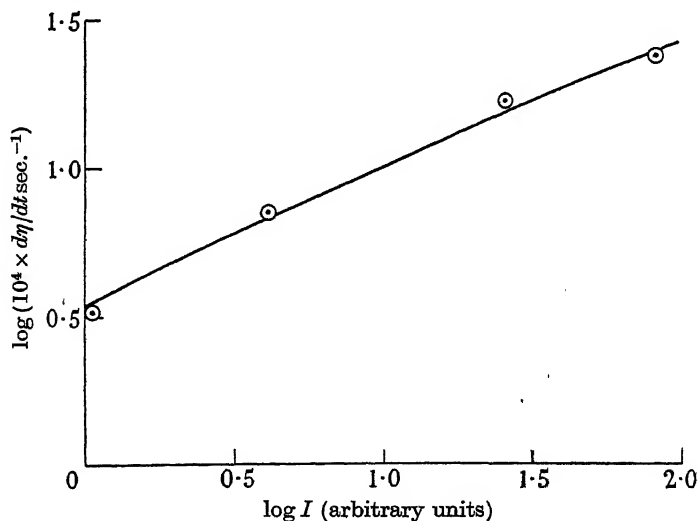
temperature 0° C		temperature -15° C	
I (arbitrary units)	$10^4 d\eta/dt$ (sec. ⁻¹) (obs.)	I (arbitrary units)	$10^4 d\eta/dt$ (sec. ⁻¹) (obs.)
43.5	8.4	1.1	3.3
80	10.8	4.1	7.1
812	36.7	26	16.9
1632	42.6	83	24.0
3182	53.5		
3223	54.5		
5753	64.0		

TABLE 3. THE RATIOS OF THE VELOCITY CONSTANTS IN THE PHOTOPOLYMERIZATION OF VINYL ACETATE

temperature (° C)	k_2/k_3	$k_3/k_4^{\frac{1}{2}}$	$A^{\frac{1}{2}}$ (arbitrary constant)
0	2.045×10^4	9.3×10^{-6}	1.66×10^{-7}
-15	2.786×10^4	4.9×10^{-6}	5.0×10^{-7}

(c) *Measurement of the photochemical after-effect*

It has already been stated that when there is no dark reaction, the photochemical after-effect should be infinite. That the observed after-effect is not infinite must therefore imply that some side reactions terminate the kinetic chains when the concentration of radicals has become very small. Although the error so introduced was unimportant when determining the photochemical rates, it becomes exceedingly important when measuring the after-effect itself. An alternative method has therefore been adopted for measuring the latter quantity. Two irradiations were carried out at the same light intensity, and for the same total time; the first irradiation being continuous, and the second intermittent, e.g. the first exposure was of 20 sec. duration, and the second was divided into four parts, each of 5 sec. duration, with 15 sec. intervals. The difference between the two viscosity increases is then equal to $3\delta\eta_I^{15}$ in equation (8).

FIGURE 3. The intensity-rate curve for the polymerization of vinyl acetate at -15°C (experimental points and calculated curve).

The light intensities were measured as before, and in the same arbitrary units. The exposure times were accurately controlled by a shutter operated electromagnetically by a clock motor and commutator.

The values of the after-effects so determined are given in table 4. The smaller after-effects at -15°C were considerably less accurately measurable than those at 0°C , and the value of the mean experimental after-effect at the former temperature is such that the termination constant would possess a small negative activation energy.

E_4 has therefore been adjusted to zero, and the value of the mean experimental after-effect at -15°C calculated using the value of $k_4(2.2 \times 10^8)$ found at 0°C . The calculated value is seen to be within the limits of experimental error.

TABLE 4. THE PHOTOCHEMICAL AFTER-EFFECT IN THE
POLYMERIZATION OF VINYL ACETATE

temperature ($^\circ\text{C}$)	I (units as in tables 2 and 3)	$\delta\eta_I^{15} \times 10^3$ (mean experi- mental value)	$\delta\eta_I^{15} \times 10^3$ (calc.)
0	5200	12.7	12.8
-15	30	6.0	5.8
		(individual experiments 6.1, 5.4, 5.1)	

RESULTS (VELOCITY CONSTANTS, ACTIVATION ENERGIES AND
FREQUENCY FACTORS)

The velocity constants (in $\text{l.mol.}^{-1}\text{sec.}^{-1}$) at 0 and -15°C for the polymerization of vinyl acetate are given in table 5. Table 6 gives the energies of activation and the frequency factors. The latter are probably correct to within a factor of 2, and the former to ± 1 kcal.

TABLE 5. THE VELOCITY CONSTANTS IN THE POLYMERIZATION
OF VINYL ACETATE

temperature. ($^\circ\text{C}$)	k_2	k_3	k_4
0	2.8×10^3	1.4×10^{-1}	2.2×10^8
-15	2.0×10^3	7.3×10^{-2}	2.2×10^8

At these temperatures the value of k_1 (thermal initiation) is so small as to be indistinguishable from zero.

TABLE 6. THE ACTIVATION ENERGIES AND FREQUENCY FACTORS IN THE
POLYMERIZATION OF VINYL ACETATE

$E_2 = 3.2$ kcal.	$A_2 = 9.8 \times 10^5$
$E_3 = 6.1$ kcal.	$A_3 = 9.9 \times 10^3$
$E_4 = 0$	$A_4 = 2.2 \times 10^8$

The quantum yield of the initiating reaction

The determination of the velocity constants described above requires only a knowledge of the relative light intensities. From equation (7) it can be seen that, if the photochemical rate corresponding to a known absolute light absorption is measured, then the quantum yield of the initiating reaction can be obtained. At very low light intensities, where the square root law is obeyed, equation (7) becomes

$$\frac{d\eta}{dt} = K' M_0^{\frac{1}{2}} [M]^2 \left(\frac{k_2}{k_3} \right)^{\frac{1}{2}} \left(\frac{k_3}{k_4} \right) (AI)^{\frac{1}{2}}.$$

The experiments were carried out in this intensity range, at a temperature of 25° C. The vacuum viscometer used had a quartz cell of 5 cm. diameter, and 5 mm. thick, fused to one end. Light from a hot mercury arc was rendered parallel by a series of quartz lenses, and was passed through a series of filter solutions as follows:

- (a) chlorine gas at 1 atm. pressure 3 cm.
- (b) solution of 50 g. $\text{NiSO}_4 \cdot 6\text{H}_2\text{O}$ in 100 ml. water 2 cm.
- (c) mixture of 40 % CCl_4 + 60 % *EtOH* by volume 5 mm.

The transmitted light consisted almost entirely of the 2650 Å line. The amounts of the neighbouring lines transmitted (2537 and 2700 to 2970 Å) were sufficiently small to be neglected in the experiments.

For measuring the incident light intensity, a uranyl oxalate actinometer was used (Leighton & Forbes 1930). It was impossible by means of a single experiment to measure both the amount of light absorbed and the photochemical rate, since the changes produced in the vinyl acetate and in the actinometer by the same amount of light energy were of quite different orders of magnitude. The light absorption by the vinyl acetate was therefore calculated from the absorption spectrum (figure 1), and the photochemical rate was measured at a given light intensity by the method previously described, with irradiation times of less than 10 min. The viscometer was then replaced by the uranyl oxalate actinometer, which was left in the light beam for 6 days. At the end of this time the change produced in the actinometer was sufficient to allow a reasonable estimate of the absolute light intensity. The actinometer cell had a thickness of 3 cm., so that the absorption of the wave-lengths falling on it could be regarded as complete.

The results of the experiments were as shown in table 7.

TABLE 7. QUANTUM YIELD EXPERIMENTS

photochemical rate	$d\eta/dt = 2.56 \times 10^{-4}$
rate of chain starting	$AI[M]^2 = 5 \times 10^{10}/\text{sec.}$
total light energy absorbed by vinyl acetate/sec.	$1.25 \times 10^{12} h\nu \begin{cases} (a) 1.35 \times 10^{12} h\nu \\ (b) 1.15 \times 10^{12} h\nu \end{cases}$
hence the quantum yield is	0.04 (± 0.01)

DISCUSSION

(i) It is of interest first to compare the values of the frequency factors and activation energies of the constituent reactions with those already known for styrene. The values for both styrene and vinyl acetate are given in table 8, the styrene figures being those obtained by Bamford & Dewar (1948a).

TABLE 8. COMPARISON OF FREQUENCY FACTORS AND ACTIVATION ENERGIES IN THE POLYMERIZATIONS OF STYRENE AND VINYL ACETATE

	frequency factors			activation energies (kcal.)	
	styrene	vinyl acetate		styrene	vinyl acetate
A_2	1.02×10^6	9.8×10^5	E_2	6.5	3.2
A_3	1.50×10^7	9.9×10^5	E_3	14.2	6.1
A_4	3.07×10^8	2.2×10^8	E_4	2.8	0

The frequency factors A_2 and A_4 are the same for both substances within the limits of experimental error, but the values of E_2 and E_4 for vinyl acetate are considerably lower than the corresponding values for styrene. This, of course, would be expected, since the resonance energy of the intermediate α -acetoxyalkyl radicals in the former polymerization is likely to be considerably lower than the resonance energy of the benzyl radicals in the latter. Some of the excess resonance energy of the benzyl radical compared with the α -acetoxyalkyl radical would undoubtedly be lost in the transition state.

The values of E_3 and A_3 for vinyl acetate are of considerable interest, since they are appreciably lower than those for styrene.

(ii) A certain amount of discussion has recently arisen regarding the values of the velocity constants in the polymerization of vinyl acetate, and particularly regarding the frequency factor for termination. Indirect support for the values given in table 6 is given by the results of Bagdassarian (1944), who measured, by a dilatometric method, the rates of polymerization at 50° C of vinyl acetate at known light intensities. The equation representing the rate of disappearance of monomer is

$$-\frac{d[M]}{dt} = \frac{k_2}{k_4} (AI)^{\frac{1}{2}} [M]^2. \quad (9)$$

Substitution in equation (9) of the values for k_2 and k_4 at 50° C calculated from table 6 allows the rate of chain starting $AI[M]^2$ to be determined. The quantum yield calculated on this basis was found to be about $\frac{1}{30}$, which is in good agreement with the present value of $\frac{1}{25}$.

Bagdassarian has also shown that the molecular weight of polyvinyl acetate prepared at low light intensities is independent of the actual value of the intensity, and that therefore the degree of polymerization is determined primarily by the chain propagation and transfer reactions. It is possible to calculate from his results an upper limit for the quantum yield assuming that such transfer does not occur, the rate of chain initiation now being obtained by means of the equation

$$-d[M]/dt = \bar{P} AI[M]^2.$$

The figure obtained was 0.28. Since the light intensities he employed were quite low, and in view of what has been said about chain transfer, it is evident that the actual quantum yield must be considerably lower than this.

In connexion with the difference $E_2 - E_3$ between the propagation and transfer activation energies, it is also interesting to observe that Bagdassarian, from his molecular weight determinations, has obtained a value of about -2 kcal. for this quantity, compared with the present value of -2.9 kcal. The low value of E_3 is therefore confirmed.

The values calculated for the propagation and termination constants at 25° C from the results of the more recent publications are given in table 9, together with the frequency factors and activation energies.

Burnett & Melville, and Swain & Bartlett both used rotating sector experiments to determine the mean life of the kinetic chains, and combined these with independent measurements of the rate of chain starting. The latter quantity was determined by

Swain & Bartlett by means of experiments in which benzoyl peroxide was used to initiate the chains. In effect, they found a concentration of benzoyl peroxide which produced a rate of polymerization equal to that observed in their photochemical experiments, and then assumed that each molecule of peroxide decomposed initiates one kinetic chain. The rates of decomposition of the peroxide at 25° C were known from previous experiments by Nozaki & Bartlett (1946). Swain & Bartlett claimed their values of k_2 and k_4 to be accurate to within a factor of 5. A simple calculation shows that, if their rate of chain initiation is reduced by a factor of 4, then the results are in excellent agreement with the present ones. In this connexion, it is interesting to note that Bamford & Dewar (1949*b*) have recently found that, in the autoxidation of tetralin initiated by benzoyl peroxide, only one chain is initiated for about ten molecules of peroxide decomposed.

TABLE 9. VALUES OF CONSTANTS IN THE POLYMERIZATION OF VINYL ACETATE

	Swain & Bartlett (1946)	Burnett & Melville (1947)	present paper
k_2 (25° C)	1.1×10^3	1.0×10^3	4.6×10^3
A_2	—	1.65×10^6	9.8×10^5
E_2	—	4.4 kcal.	3.2 kcal.
k_4 (25° C)	8×10^7	3×10^8	2.2×10^8
A_4	—	3×10^9	2.2×10^8
E_4	—	0	0

Burnett & Melville, on the other hand, determined the rate of chain initiation by measuring the rate of disappearance of inhibitor (*p*-benzoquinone) at a known light intensity. It was assumed that one molecule of the quinone was removed per kinetic chain, and that all chain initiation was due to light absorption by the vinyl acetate. However, as has already been pointed out by the author (1947), the absorption spectrum of *p*-benzoquinone is such that there is no spectral region where the inhibitor technique using this substance can be used with pure vinyl acetate, due to its internal filter effect. The argument that the duration of the induction period of the polymerization is directly proportional to the amount of inhibitor is not convincing, since so little is in fact known about the details of the action of quinone inhibition. Burnett & Melville's value of unity for the quantum yield of the initiating reaction must therefore be regarded as unreliable, particularly as it is considerably higher than the maximum possible value calculated earlier.

Even assuming a more reliable value for the quantum yield it is impossible to bring Burnett & Melville's results into agreement with those of Swain & Bartlett or those of the present paper. In fact, on reducing the value of the quantum yield from unity, the discrepancy in the values of k_4 becomes still larger. However, it is difficult to discover an exact reason for the disagreement since Burnett & Melville failed to state their exact experimental conditions. Professor Melville, in a private communication to Swain & Bartlett (see Swain & Bartlett 1946, p. 2386) has stated that light was used in the sector experiments of wave-length greater than 2500 Å, and this would agree with the statement that a cold acetic acid filter was employed. Swain & Bartlett have therefore suggested that the discrepancy is due to almost

the whole of the light being absorbed by the vinyl acetate in an extremely small portion of the reaction cell; and approximate calculations show that this certainly brings the results into much better agreement. The high concentration of growing chains in such a portion of the cell would also explain why the chain transfer reaction could not be detected in their experiments. However, the explanation would appear to be contradicted by the statement of Melville & Burnett (1947) that, under the conditions which they originally used, a 40 % solution of the monomer in *n*-hexane gives practically the same values for the propagation and termination constants as does the pure monomer.

It is also possible that Burnett & Melville's method of purification of the vinyl acetate was inadequate, and this would, in addition, explain the differences between the absorption spectra in figure 1.

In conclusion, it may be said that the bulk of the experimental evidence seems to support the values for the velocity constants given in the present paper.

REFERENCES

- Bagdassarian, Ch. 1944 *Acta Physicochim. U.R.S.S.* **19**, 266.
Bamford, C. H. & Dewar, M. J. S. 1948*a* *Proc. Roy. Soc. A*, **192**, 309.
Bamford, C. H. & Dewar, M. J. S. 1948*b* *Proc. Roy. Soc. A*, **192**, 329.
Bamford, C. H. & Dewar, M. J. S. 1949*a* *Proc. Roy. Soc. A*, **197**, 356.
Bamford, C. H. & Dewar, M. J. S. 1949*b* *Proc. Roy. Soc. A*, **198**, 252.
Briers, F., Chapman, D. L. & Walters, E. 1926 *J. Chem. Soc.* p. 562.
Burnett, G. M. & Melville, H. W. 1947 *Proc. Roy. Soc. A*, **189**, 456.
Dixon-Lewis, G. 1947 *Disc. Faraday Soc.* **2**, 319.
Houwink, R. 1940 *J. prakt. Chem.* **157**, 15.
Leighton, W. G. & Forbes, G. S. 1930 *J. Amer. Chem. Soc.* **52**, 3139.
Melville, H. W. & Burnett, G. M. 1947 *Disc. Faraday Soc.* **2**, 370.
Nozaki, K. & Bartlett, P. D. 1946 *J. Amer. Chem. Soc.* **68**, 2377.
Schou, S. A. 1929 *J. Chim. phys.* **26**, 77.
Schulz, G. V. & Sing, G. 1943 *J. prakt. Chem.* **161**, 161.
Swain, C. G. & Bartlett, P. D. 1946 *J. Amer. Chem. Soc.* **68**, 2381.
Wagner, R. H. 1947 *J. Polymer Sci.* **2**, 21.

The ignition of solid explosive media by hot wires

By E. JONES, *I.C.I. Limited, Nobel Division, Stevenston, Ayrshire*

(Communicated by F. P. Bowden, *F.R.S.*—Received 29 January 1949—
Revised 7 April 1949)

The present paper describes an investigation of the physical factors affecting the ignition of solid explosives by heated filaments embedded in the medium. The filaments were composed of fine resistance wire and were heated electrically, the critical thermal energy required to cause ignition being measured for wires of different geometrical, thermal and electrical characteristics and for different times of heating. Systematic variation of these factors enabled the energy equation for the ignition process to be formulated and its terms analyzed, the technique involving extrapolation to zero time of heating as a means of eliminating heat losses from the ignition system, and extrapolation to zero diameter of wire in order to eliminate terms involving the heating element; the former simulates the ideal case of a heat-insulated ignition system and the latter that of a line source of heat.

The energy equation for ignition in these circumstances takes a simple form which implies that, at the moment of ignition, the heat supplied to the ignition system always equals the heat gained by the system plus the heat lost, the absence of any term representing heat generated by chemical action being very significant. For a given ignition system, the amount of heat absorbed up to the moment of ignition is shown to be independent of time, so that the increase in ignition energy with increasing time of ignition is wholly attributable to the heat losses sustained by the ignition system during the heating process. Further analysis shows that the critical factor governing ignition in systems of the type considered is the temperature, and that the geometry of the heating element probably determines the amount of explosive which must be raised to the critical temperature to ensure ignition.

INTRODUCTION

A thermal theory of ignition and flame propagation is attractive, especially to physicists and engineers, because it involves temperature and heat, the essential elements for the application of thermodynamics to explosion problems. Briefly, the thermal theory postulates that an explosive medium ignites spontaneously when its temperature reaches a certain critical value, known as the 'ignition temperature', and that a flame, once started, is capable of self-sustained propagation if the heat communicated from the burning layer to the adjacent unburnt layer is sufficient to raise the latter to the ignition temperature. One method of testing this theory would be to generate heat in the explosive medium and deduce the critical temperature from the critical quantity of heat required for ignition. Care should be taken, however, to choose an unambiguous source of heat, since the generation of heat, especially in a gas, is liable to cause expansion and, if this occurs very rapidly, much of the energy of the source may be wasted in doing mechanical work. For example, an electric spark passed through a loose heap of even sensitive explosives like gun-cotton or mercury fulminate may succeed only in scattering the material, whereas a smaller spark will cause ignition if the explosive is constrained. A similar effect may be observed when a detonator is fired in a heap of gunpowder or when a charge of certain high explosives is fired in an explosive firedamp atmosphere. It is not sufficient merely to measure the energy dissipated by the igniting source; the proportion of this energy appearing as sensible heat in the medium must be determined.

Again, if the process of raising the explosive to its ignition state occupies a finite time, some loss of heat to the surroundings is inevitable and such losses should, of course, be taken into account.

Bearing these observations in mind, it will be seen that the use of a solid medium, with its negligibly small coefficient of expansion in comparison with that of a gas, offers a means of reducing to insignificance the energy wasted in doing work against atmospheric pressure, no matter how rapid the heating process. Further, the use of a rigid medium inhibits relative movement, thus preventing loss of heat by convection or by creating kinetic energy. These advantages appear to be afforded by the method described by Morgan (1925), wherein use was made of the small electric igniters, or 'low-tension fuseheads', which form a part of the ordinary commercial electric detonator.

The low-tension electric fusehead, as used by Morgan, consists essentially of a match-head formed round a small resistance element, the heating of which on passing a suitable current causes the match-head to ignite. The match-stalk consists of a narrow strip of cardboard with metal foil on either side, the foils being 'bridged' at the tip by a fine resistance wire attached by soldering. The tip is 'stepped' so that when the match-head is formed by the usual process of 'dipping', the sensitive composition fills the step and thus completely surrounds the bridge-wire for most of its length. The construction of the device is shown diagrammatically in figure 1.

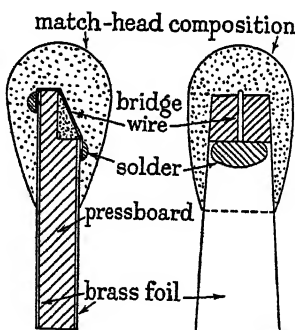


FIGURE 1. Diagrammatic sketch of low-tension fusehead.

In Morgan's experiment 'the object was to find the current required for ignition when the duration of current was limited to a definite interval of time', and the conclusion was that 'experiments on the ignition of highly inflammable solids by means of very short hot wires show that the energy required for ignition increases with diminution of the rate of heating and that, over a wide range, a linear relation exists between the energy and the time during which it is supplied'.

If the assumption is now made that the ignition energy has a specific value for a given type of fusehead, it follows that its apparent increase with diminution of the rate of heating must be due to increasing energy losses with time; in which case the losses can conveniently be eliminated by extrapolating to zero time of heating. This critical value can be derived from Morgan's results, but its full significance cannot be ascertained because of the uncertainty regarding the way in which this energy is

divided between the bridge-wire and the explosive medium. If by a suitable extension of Morgan's experiment this difficulty can be resolved, it should then be possible to measure the critical energy which must be given to the explosive in order to effect ignition. The present experiments were undertaken with this object in view.

It is significant to note that the above argument is based on the assumption that, in ignition systems of the type under consideration, the critical energy for ignition is independent of time and that it is the loss of energy in the ignition process which varies with time. This, therefore, is the hypothesis underlying the present investigation and one object of the work will be to test this hypothesis. It is also worth noting that this hypothesis is implicit in our statement of the thermal theory of ignition so that this test is, in fact, a test of the thermal theory.

EXPERIMENTAL

For commercial use, fuseheads of the type described require the use of very sensitive explosives, but, so that the present investigation should not be unduly restricted in this respect, three match-head compositions of varying sensitivity were examined. The compositions were chosen to include both single explosive compounds and heterogeneous explosive mixtures, one a mechanical mixture of non-explosive ingredients. By thus introducing a wide range of chemical variables, it was hoped to subject our purely physical hypothesis to a more stringent test. The compositions actually selected were: (1) copper acetylide, (2) a mixture of four parts by weight of lead mononitroresorcinate (LMNR) to one of potassium chlorate, and (3) a mixture of five parts by weight of potassium chlorate to one of birchwood charcoal. The powders were bonded together to form a rigid matrix by the use of a small amount of nitrocellulose in each case.

The bridge-wire can vary in material, length and diameter, all of which are capable of more or less rigid control. The bridges actually employed in the experiments comprised a series of nichrome wires varying in length from 0.9 to 2.9 mm. and in diameter from 1.3 to 7.8×10^{-3} cm., together with some miscellaneous wires of copper, platinum, tin and lead.

To fire the fuseheads, direct current was applied through a pendulum time-switch which could be adjusted to give time intervals from 1 to 60×10^{-3} sec. The instrument was calibrated at each setting by measuring with a fluxmeter the quantity of electricity passing through the switch on applying a known small current.

In each experiment carried out at a particular setting of the time-switch, some preliminary trials were made to ascertain roughly the value of the mean firing current corresponding to the time of application concerned and then twenty tests, each with a fresh fusehead, were made at each of two currents, one slightly above and the other just below the expected mean, the proportion of ignitions occurring at each current being noted. The true value of the mean firing current for that particular time of current application was obtained by interpolating the point at which 50 % ignitions would occur.

It may be noted that the 'mean firing current' is associated with a unique value of the application time and, to avoid any ambiguity, this will be called the 'excitation time'. In other words, the excitation time is the time for which a specified current must be applied in order to fire, on the average, 50 % of the fuseheads. Similarly, the 'mean firing current' is the value of the firing current which, when applied for a stipulated time, causes 50 % of the fuseheads to ignite.

Time of current application

Using the experimental procedure outlined above, the mean firing currents corresponding to various times of current application were determined for a batch of fuseheads of the same kind. In this case, the bridge-wire had a mean length of 1.51 mm. and a mean resistance of 1.18Ω , the original wire being nichrome of diameter 0.0042 cm. and linear resistance $7.8\Omega/\text{cm}$. The fusehead composition consisted of four parts of LMNR to one of potassium chlorate, the materials being bonded together with nitrocellulose. Four periods of current application were selected, viz. 50, 25, 12 and 5×10^{-3} sec., and the results obtained are represented graphically in figure 2.

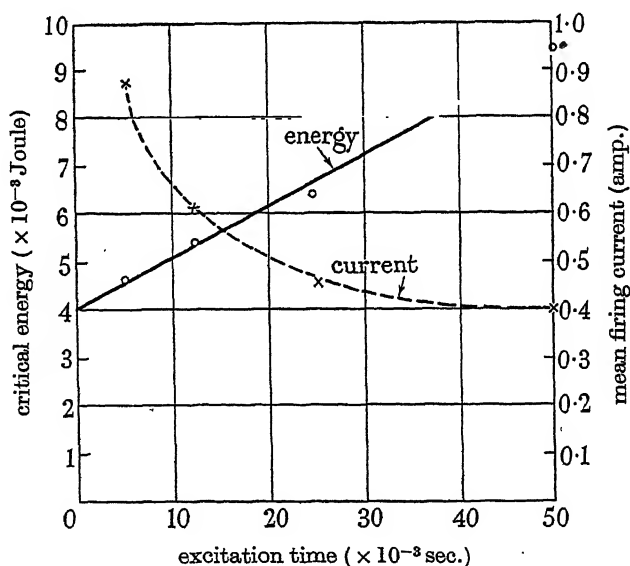


FIGURE 2. Relation between ignition energy and excitation time.

It will be seen that, whereas the mean firing current falls with increasing application time, the critical energy rises and, moreover, that a linear relationship exists between the critical energy for ignition and the excitation time, confirming Morgan's earlier conclusion. Thus, the results, expressed in terms of energy, can be represented by the following equation:

$$E = A + Bt, \quad (1)$$

where E is the critical energy for ignition corresponding to an excitation time t , A is the intercept on the energy axis and B is the slope of the line. So far, then, the results confirm our hypothesis in that the critical energy for ignition under given

conditions comprises two parts, one independent of time and the other a function of time. The results further show that the latter term is directly proportional to the excitation time.

Equation (1) is the energy equation for the ignition process and, in the general case, will contain thermochemical and thermodynamic as well as sensible heat terms. Since, however, the means of ignition in these experiments was purposely chosen to avoid thermodynamic complications, this equation may be taken to express a heat balance, in which case it may be interpreted as meaning that, of the total heat, E , supplied to the ignition system, a part, A , is retained in the system and the remainder, Bt , escapes. Accepting this interpretation, it follows that B represents the rate of loss of heat by the system, and the interesting conclusion is reached that this is a constant with respect to excitation time.

It is desirable at this point to form some picture of the ignition system, and it seems reasonable to suppose that the ignition system comprises the bridge-wire and that portion of the surrounding explosive medium which must be brought to some critical condition in order that a self-sustained reaction becomes possible. As yet, the amount of explosive involved in an ignition is unknown, but if, as a first approximation, it is assumed that all the heat retained by the system is contained in the wire, the temperature of the wire can be calculated, and this, at least, gives a figure which the temperature of the system as a whole cannot exceed. The value of A obtained from figure 2 is approximately 4.1×10^{-3} joule, or 1×10^{-3} cal., and, as the thermal capacity of the wire was about 2×10^{-6} cal./°C, the maximum wire temperature is approximately 500°C, which, as we have seen, must be an over-estimate. On the other hand, the rate of loss of heat, B , is approximately 25×10^{-3} cal./sec., so that, if the excitation time—or, as it might be called, the 'lag on ignition'—were 1 sec., the quantity of heat required to effect ignition would be 26×10^{-3} cal., i.e. sufficient to raise the wire to a temperature of about 13,000°C. It will be seen, therefore, that when the ignition process occupies a finite time and due allowance is not made for heat losses, the apparent 'ignition temperature' is liable to be grossly exaggerated.

As the derivation of the quantity A involves extrapolation to zero excitation time, the physical significance of this procedure requires examination. The times measured are, of course, the times of application of the heating current and, in extrapolating to zero excitation time, we are simulating conditions in which sufficient heat is supplied instantaneously to make ignition inevitable but not necessarily immediate. The instantaneous supply of a finite quantity of heat to a real system is, of course, a physical impossibility, and, even if such a thing were possible, it is not likely that uniform distribution of the heat throughout the system would result immediately. On the other hand, if these processes occupy a time which is short relative to the measured excitation times, they can be assumed to occur instantaneously, relatively speaking. A clearer picture is probably afforded if the extrapolation process is regarded as a convenient method of approaching the ideal state in which no heat is lost from the ignition system, i.e. the ideal case of a perfectly insulated system. The rate of loss of heat from the system is given by the slope of the line, and the constancy of this slope throughout the experimental range justifies extrapolation to the

hypothetical point where the time interval is too short to permit any such heat to escape. The heat contained in the system at this point may be expressed in calories or, alternatively, as a function of the temperature of either the wire itself or the ignition system as a whole.

Bridge-wire characteristics

Since the ignition system comprises only two parts, it follows that, if the total energy retained in the system is known, the amounts claimed by both components are known once the portion contained in any one component is isolated. It has already been seen that extrapolation to zero time simulates the case of a heat-insulated system, which satisfies our first condition. With a wire of given material, e.g. nichrome, the bridge-wire is completely characterized by its length and diameter, so that, if these two factors are successively eliminated, the wire vanishes and we are left with a one-component system, thus satisfying our second condition.

This process, however, gives no direct information regarding the effect of the characteristics of the wire material and, although the relevant properties must be self-evident once we postulate a purely thermal energy equation, the correctness, or otherwise, of this assumption can be subjected to the test of experiment by using wires of different thermal properties. Hence, it was decided to determine the effects on the critical energy for ignition of (a) length of bridge-wire, (b) diameter of bridge-wire, and (c) material of bridge-wire.

(a) Bridge-wire length

To examine the effect of bridge-wire length on the critical energy for ignition, five batches of fuseheads were made in which the 'step' lengths were varied so as to alter the lengths of the bridge-wires. The fuseheads were bridged, as before, with nichrome wire of diameter 0.0042 cm. and linear resistance $7.8\Omega/\text{cm}$. As it was necessary to test the electrical resistance of each fusehead, the most convenient method of determining the lengths of the bridge-wires was to calculate them from the fusehead resistances and the known linear resistance of the wire. The mean lengths of the bridge-wires in the five lots of fuseheads were 0.87, 1.03, 1.51, 2.03 and 2.85 mm. respectively. The same match-head composition was used throughout, namely, 80 % of LMNR and 20 % of potassium chlorate. As before, the mean firing currents corresponding to four different application times, ranging from 5 to 50×10^{-3} sec., were determined for each of the five batches, the results of this experiment being represented graphically in figure 3.

As before, the results are represented by straight lines, the critical energy decreasing with decreasing excitation time but remaining finite when extrapolated to zero excitation time. Both the intercepts on the energy axis and the slopes of the lines vary with the length of the bridge-wire, so that, in the light of this new evidence, our original energy equation must be qualified, since it is true only for a particular length of bridge-wire. Consequently, we must write

$$E = A + Bt, \text{ when } l \text{ is constant,} \quad (2)$$

where l is the length of the bridge-wire.

In figure 4, the critical energies for ignition have been plotted against length of bridge-wire for different excitation times, and, again, the results can be represented

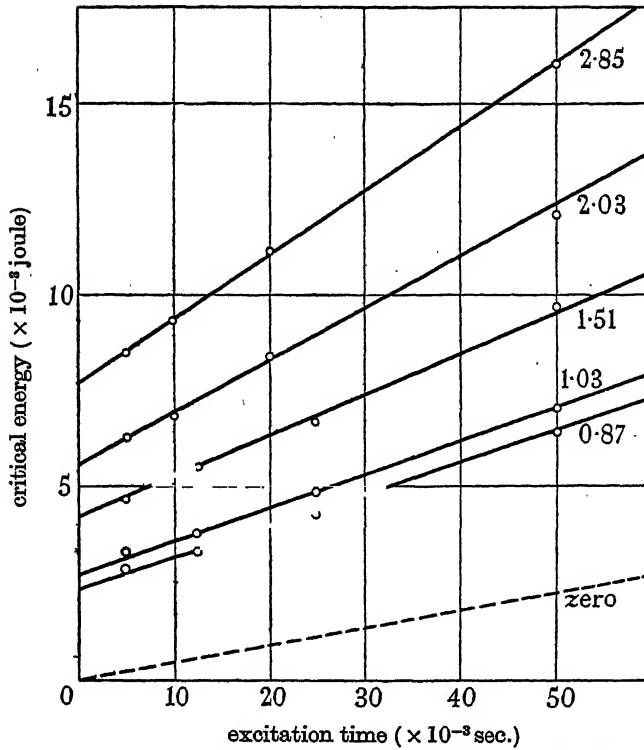


FIGURE 3. Effect of time of current application on critical energy for ignition. (Bridge-wires of different lengths shown in mm.)

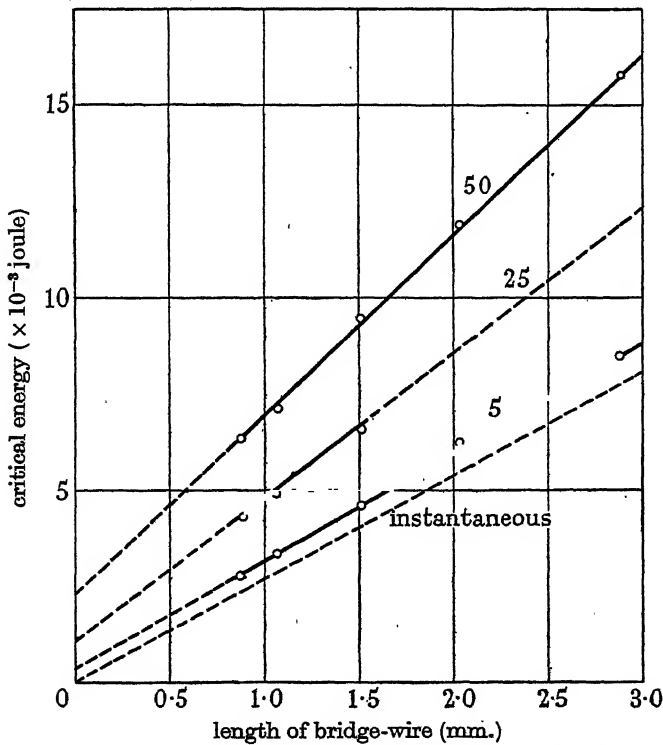


FIGURE 4. Effect of bridge-wire length on critical energy for ignition (varying excitation time shown in sec. $\times 10^{-3}$).

by a series of straight lines cutting the energy axis at positive values and having positive gradients. Both the intercepts on the energy axis and the slopes of the lines vary depending on the excitation time, so that the relationship between critical energy for ignition and length of bridge-wire may be represented by

$$E = C + Dl, \text{ when } t \text{ is constant,} \quad (3)$$

and, on combining (2) and (3) algebraically, we have

$$E = F + Gl + Ht + Jlt, \text{ when both } l \text{ and } t \text{ vary.} \quad (4)$$

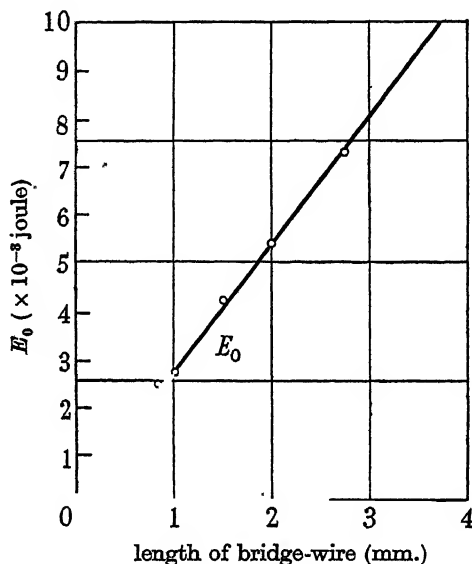


FIGURE 5. Effect of bridge-wire length on E_0 .

The results, however, show that when the intercepts, E_0 , on the energy axis in figure 3 are plotted against length of bridge-wire, as in figure 5, the line passes through the origin, which signifies that, when $t = 0$, equation (4) reduces to $E = Gl$, and, therefore, $F = 0$. Our energy equation thus becomes

$$E = Gl + (H + Jt)t. \quad (5)$$

The physical interpretation of equation (5) is still that the heat supplied to the ignition system is equal to the heat absorbed plus the heat lost, but now we find that the heat absorbed is directly proportional to the length of the bridge-wire, whereas the heat losses are of two kinds, one independent of the length of the bridge-wire and the other directly proportional to it. As the only variable concerned is the linear dimension of the bridge-wire, the fact that the heat losses are of two kinds immediately suggests end-effects and lateral effects. In other words, there is loss of heat from the ends to the soldered joints and the metal foils, which is independent of the length of the bridge-wire, and there is also lateral loss of heat through the explosive medium, this being directly proportional to the length of the bridge-wire. The results are, therefore, consistent with our original supposition that the term Bt in equation (1) does, in fact, represent the heat losses. Further, we have now arrived

at the interesting conclusion that the heat transmitted to the explosive medium per unit length of wire is the same at all points along the wire.

Since the first term on the right-hand side of equation (5) represents the heat retained in the system, we can calculate, as before, the temperature of the wire on the assumption that the whole of this heat remains in the wire. This was done earlier for one length of bridge-wire, i.e. 1.51 mm., and, since E/l is constant when $t = 0$, the temperature will be the same for all lengths of bridge-wire, namely, approximately 500° C. In other words, the temperature is constant all along the length of the bridge-wire, and it is not surprising that both the rate of loss of heat at the ends and the rate of loss of heat into the composition per unit length of wire are constant.

(b) *Bridge-wire diameter*

The first series of experiments made with a view to examining the effect of bridge-wire diameter on the critical energy for ignition was carried out with a range of nichrome wires varying in diameter from 0.00126 to 0.00422 cm. The step length was kept constant with the object of preserving a constant bridge-wire length throughout the series, but, owing to difficulties in the manipulation of such fine and fragile wires, control of this and other experimental factors was not so effective as in the earlier experiments. One batch of fuseheads was made with each diameter of wire, the mean length of the bridge-wire in each case being estimated, as before, from the mean fusehead resistance and the known linear resistance of the original wire. The match-head composition was the same as before, i.e. 80 % of LMNR and 20 % of potassium chlorate, and the fuseheads were tested along similar lines.

It will be seen from figure 6, where the critical energy per cm. length of bridge-wire has been plotted against excitation time for each diameter of bridge-wire, that the characteristic linear relationship is again evident, but the family of lines is not so well ordered as in the previous experiment, presumably due to uncontrolled experimental variables. The inconsistency appears to be associated with the slopes of the lines, with the consequence that the discrepancies are magnified at the higher excitation times. On extrapolating to zero excitation time, the critical energies are found to be arranged in a logical order, increasing progressively with increasing diameter of bridge-wire. These features accord with our hypothesis, which predicts that experimental conditions may well affect the leakage of energy, but the basic criteria for ignition by a standard method, or model, are characteristic of the explosive medium, which, of course, was constant throughout the series. However, the analysis of the quantity G requires a knowledge only of the critical energies at zero excitation time, and the appropriate values for the different diameters of bridge-wire, obtained from figure 6, are plotted against the area of cross-section of the bridge-wire in figure 7. It will be seen that the results can be represented by a straight line, the equation of which may be written

$$G = E_0/l = K + La, \quad (6)$$

where E_0 is the critical energy corresponding to zero excitation time, a is the cross-sectional area of the bridge-wire and K and L are constants with respect to a .

Thus the term G is shown to consist of two parts, one independent of the geometry of the heating element and the other directly proportional to the cross-sectional area

of the bridge-wire. The inference is that we have at last separated the two components of our ignition system, the term La referring to the bridge-wire and K to the portion of the explosive medium actively involved in an ignition. Accepting this view and bringing equation (6) back into energy units, we have

$$E_0 = Kl + Lal = Kl + Lv, \quad (7)$$

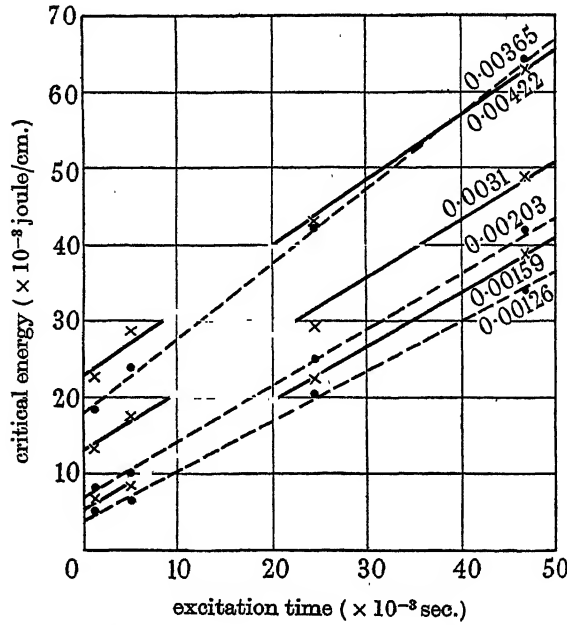


FIGURE 6. Relation between excitation time and critical energy for ignition. (Bridge-wires of different diameters shown in cm.)

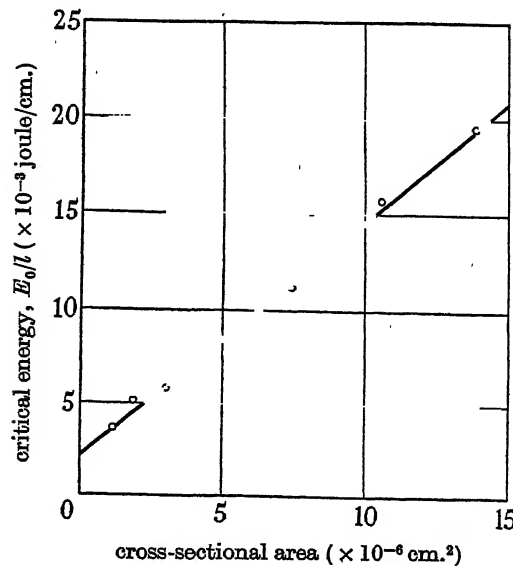


FIGURE 7. Effect of area of cross-section of bridge-wire on E_0/l .

where v is the volume of the bridge-wire. This, of course, is the energy equation for a completely insulated ignition system, and we now see that the energy actually used in igniting the explosive medium is Kl which, as might have been expected, is directly proportional to the length of the ignition system. The heat contained in the bridge-wire at ignition, and so not effectively participating in the ignition of the explosive medium, is, on the contrary, directly proportional to the volume of the bridge-wire. Moreover, the constant L has the dimensions of heat per unit volume and, with wires of the same material, i.e. of the same thermal capacity per unit volume, this means that the temperature of the wire at ignition is constant irrespective of the length or diameter of the bridge-wire. In other words, we have not yet succeeded in making any change in the bridge-wire which has had any effect on the temperature obtaining at ignition. We are thus led to conclude that this temperature is not in any way affected by the characteristics of the bridge-wire and must, therefore, be a characteristic of the explosive. In other words, the 'ignition temperature', as postulated by the thermal theory of ignition, is a physical reality.

Having apparently isolated the heat contained in the bridge-wire, it is now possible to make a closer estimate of its temperature, i.e. the temperature at which it ignites a mixture of four parts of LMNR to one part of potassium chlorate bonded together with nitrocellulose. We obtain L from the slope of the line in figure 7, and this is approximately 1450 joule/cm.³, or about 350 cal./cm.³. The thermal capacity of nichrome being approximately 0.92 cal./cm.³/°C, the temperature of the wire at ignition would appear to be about 380° C. This is only a rough estimate, which ignores such factors as the temperature coefficients of resistance and specific heat, but as the present discussion is primarily concerned with the principles of ignition, numerical figures are of secondary importance. The significant point is that, whereas our first approximation gave a value of 500° C, our second gives 380° C, and we may further anticipate that, when the energy equation is fully established, this too may have to be amended.

(c) *Bridge-wire material*

The results have now led to the following formulation of the energy equation for ignition of a solid explosive medium by hot wires:

$$E = (Kl + Lv) + (H + Jl)t = E_0 + kt, \quad (8)$$

$$\text{where} \quad E_0 = Kl + Lv \quad (7)$$

$$\text{and} \quad k = H + Jl. \quad (9)$$

Equation (9) gives the rate of dissipation of heat outside the ignition system. No attempt will be made here to analyze this factor, largely because it plays a secondary role in the ignition process and its consideration in detail would complicate the discussion unnecessarily. It may play a more prominent role in the propagation of ignition, that is, if, as is commonly supposed, the unignited material is brought to its ignition point by heat conduction.

Taking equation (7), we have deduced that the term Kl refers to the explosive component of our ignition system and, for a given length of bridge, may therefore

be expected to be indifferent to variations in the material of the bridge-wire. Hence, dividing throughout by l , we have

$$E_0/l = K + Lv/l,$$

where K is now a constant with respect to bridge-wire material. Also, according to our interpretation, the term Lv/l represents the heat retained in unit length of bridge-wire and equals the rise in temperature, T , multiplied by the thermal capacity, c , of the wire per unit length. But we have deduced that, for a given explosive medium, T is constant, so that, if we vary the bridge-wire material, keeping the explosive medium constant, equation (7) becomes

$$E_0/l = K + Tc, \quad (10)$$

where T is now a constant. Thus, by plotting E_0/l against c , we should, if our inferences are correct, obtain a straight line, the inclination of which is the actual rise in temperature of the wire.

Strictly speaking, these predictions are based on the assumption that the geometry of the ignition system is not significantly altered, but it is not easy in practice to preserve the same dimensions of bridge-wire and still provide a reasonable range of variation in thermal properties. Nichrome, platinum and copper wires of comparable diameters were conveniently available, but the lead and tin wires used were much thicker, the former being approximately seven times and the latter about four times the diameter of the first three. These materials afforded a wide range of specific resistance, specific gravity, specific heat and thermal conductivity, besides which it was hoped that the low melting-points of tin and lead, at which fusion of the wires might be expected, would provide an independent check on the temperature. Using these wires to form the bridges and 80/20 LMNR/potassium chlorate as the match-head composition, a series of fuseheads was made and submitted to the usual tests. Some difficulty was experienced in manipulating the lead and tin wires so that a longer bridge had to be accepted in these two cases, the bridge-wire length being 7.7 to 7.8 mm. compared with 1.4 to 1.5 mm. for nichrome, platinum and copper. For each type of fusehead, the mean firing currents at four or five different application times were determined, but the range of application times obtainable with the tin and lead wires had to be limited because the fuseheads required firing currents greatly exceeding any contemplated when the apparatus was designed. The results obtained are represented graphically in figure 8, where the critical energies per unit length of bridge-wire have been plotted against the excitation time.

Except for the lead wire, which required very heavy firing currents and gave rather erratic results, the characteristic straight-line relationship between critical energy for ignition and excitation time is again reproduced. In the case of the lead wire, the line has been drawn in what is considered the most likely position and can only be regarded as a plausible representation of the experimental results.

The values of E_0/l , expressed in cal./cm., have been plotted against the thermal capacity of the wire/cm. length in figure 9, and it will be seen that the results suggest a linear relationship between these two quantities, as predicted. The slope of this line, which should represent the temperature of the wire at ignition, gives a numerical

value for the uncorrected wire temperature of approximately 310°C , which compares reasonably well with the previous figure of 380°C considering the magnitude of the quantities of heat these measurements involve. Again, by extrapolating to zero heat

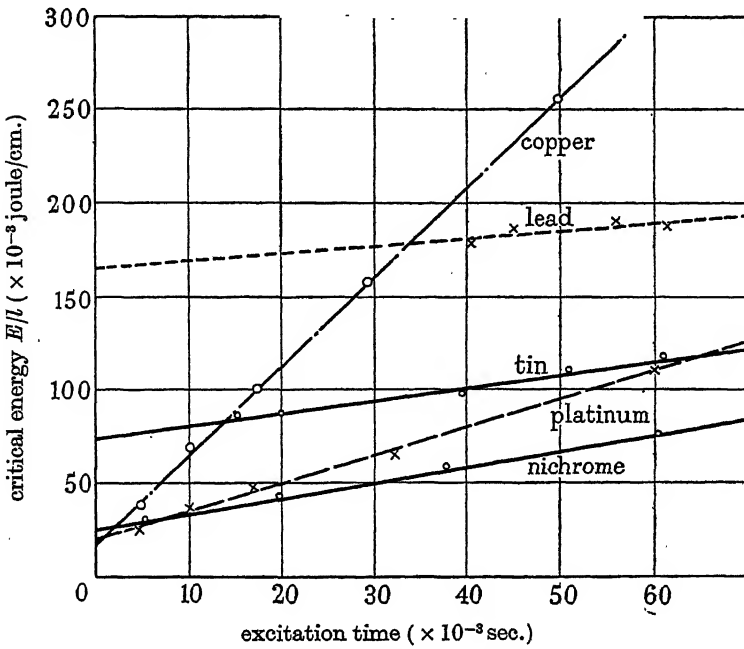


FIGURE 8. Effect of bridge-wire material on critical energy for ignition.

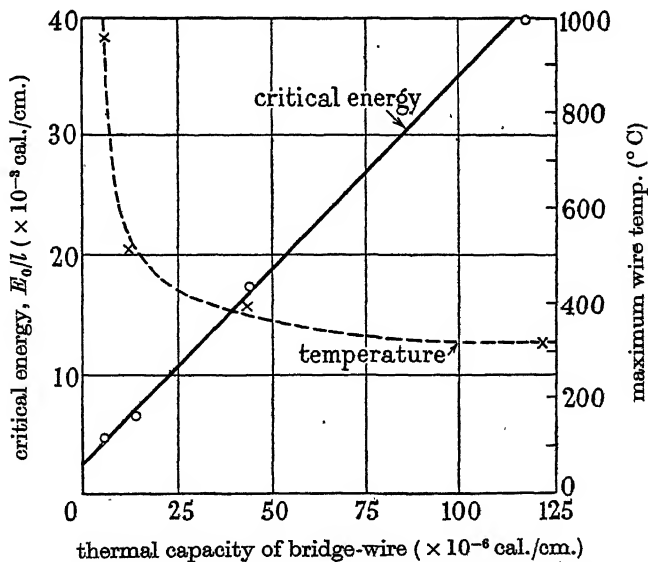


FIGURE 9. Relation between critical energy for ignition and thermal capacity of bridge-wire.

capacity, we can eliminate the substance of the wire—but not its diameter on this occasion—and find how much heat is required to ignite the explosive medium when none is retained by the wire. The numerical value of this critical quantity of heat is

approximately 2.5×10^{-3} cal., which is of the same order of magnitude as the previous estimate of 0.5×10^{-3} cal. It is difficult to say at this stage whether or not the discrepancies between these numerical estimates are significant, but the fact that an increase in the quantity of heat attributed to the explosive medium is accompanied by a drop in the temperature of the wire is rather disturbing. This point will be reverted to later.

Another interesting test which can be applied to the results is to calculate, as before, the temperature of the bridge-wire on the assumption that no heat is given up to the explosive medium. The results of this calculation are represented in figure 9, and it will be seen that, with increasing heat capacity of the bridge-wire, the apparent temperature at first falls extremely rapidly but soon approaches a steady value which agrees with the figure given above, namely, approximately 310° C. Thus, extrapolation to infinite heat capacity gives much the same result as extrapolating to zero heat capacity, which further emphasizes the fact that the temperature of the wire at ignition has little to do with the characteristics of the bridge-wire.

The attempt to estimate the ignition temperature from the melting-points of the wire materials, on the assumption that a wire fuses at or near its melting-point and, by breaking the circuit, arrests any further rise in temperature, was not successful, possibly because the molten metal was retained in position by the rigid matrix surrounding it. On the other hand, the possibility that the actual temperature of ignition may be much lower than these approximate estimates cannot be entirely ignored.

Fusehead composition

It has been shown that, to effect ignition of a solid explosive medium by a heated filament, it is necessary not only to establish a certain critical temperature but also to communicate a certain critical amount of heat. Whereas the first criterion follows from the thermal theory, the significance of the latter is not immediately obvious. If, however, we assume that the temperature of the explosive component of our ignition system is constant, the critical quantity of heat corresponds to a critical quantity of explosive heated to the critical temperature, and we have now introduced a geometrical factor.

The ignition system has finite dimensions and must be expected to remain finite even when the heating element tends towards vanishing point. At this point the system has been reduced to its most elementary form; the source of heat is a mathematical line having no substance, and the ignition system contains explosive medium only. If we now make the bold assumption that this limiting size is determined by the geometry of our chosen model, i.e. that it has no connexion with the thermochemistry of ignition, we can treat this factor as a constant with respect to the explosive medium. Interpreted broadly, this means that, with a line source of heat, the amount of explosive concerned in an ignition is the same whatever explosive is used, and the critical quantity of heat will be proportional to the critical temperature for all explosive media. While this may be far too simple a view of the ignition process, it provides a basis for analyzing the relationship, if any, between these two critical quantities when the explosive medium varies.

The explosive media used for this purpose were (a) copper acetylide, and (b) a mixture of five parts by weight of potassium chlorate to one of birchwood charcoal, the solids being bonded together, as before, with nitrocellulose. The preparation and testing of the fuseheads followed the same lines as those described previously. The critical energies at zero excitation time, expressed in cal./cm., have been plotted against the thermal capacity of the bridge-wire, also expressed in cal./cm., in figure 10 for the two compositions concerned. This graph also includes the earlier data

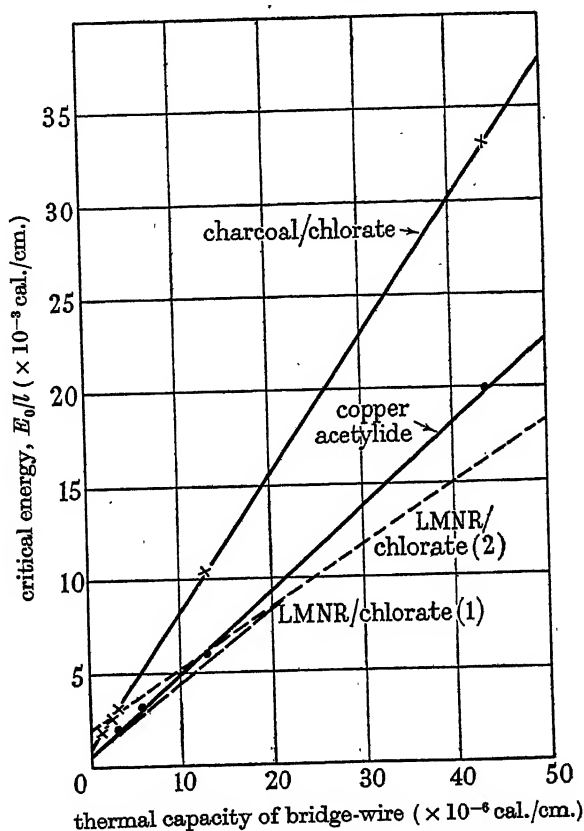


FIGURE 10. Relation between critical energy for ignition and thermal capacity of bridge-wire.

for the LMNR/chlorate composition obtained (1) by varying bridge-wire diameter, and (2) by varying bridge-wire material. Figure 10 thus summarizes the results obtained in the present investigation.

As before, the relationship between the critical energy for ignition and the thermal capacity of the bridge-wire is linear for a given explosive, but the results now show that this relationship varies from one explosive to another. From the fact that the slope of the line was found to be independent of the characteristics of the bridge-wire, we had already deduced that it was probably a function of the explosive medium, and it is now shown that this inference was correct. Again, the slope of this line is a rough measure of the apparent temperature of the bridge-wire at ignition and, confining our attention to the results obtained by extrapolation to zero wire diameter, we find that the critical temperatures for the LMNR/chlorate, copper acety-

lide and charcoal/chlorate compositions are respectively 380, 440 and 740° C approximately. Associated with these critical temperatures we find critical quantities of heat of approximately 0.5, 0.6 and 1.0×10^{-3} cal./cm. respectively. It will be noted that these two criteria for ignition bear an approximately constant numerical ratio one to the other and, while some doubt is felt if the accuracy of the present method justifies this extreme step in the analysis, this does suggest, as we have seen, that with a line source of heat, the amount of explosive concerned in an ignition is constant. If the temperature depends on the explosive and the quantity on the source, then it may be inferred that, in the ignition of an explosive medium by a source of heat conforming to a particular geometrical model, certain geometrical requirements are superimposed on those arising from the physico-chemical properties of the medium itself.

Referring to the results obtained with bridge-wires of different materials, it will be appreciated that, in this experiment, the substance of the wire was made to vanish without eliminating its size and shape, the result being that our ignition system was left with an empty space along its axis. According to our results, this had the effect of slightly reducing the critical temperature and substantially increasing the critical quantity of explosive concerned in an ignition. This suggests that the critical amount of explosive tends to increase with increasing diameter of the heating element, and this obviously throws some doubt on the validity of our assumption that the whole of the energy represented by the term Lv is contained in the bridge-wire. If part of this energy is actually located in the explosive medium, our estimated temperatures are still too high, and, until this point is clarified, the calculated temperatures can only be regarded as approximations, the true temperatures being probably somewhat lower.

DISCUSSION

Taking the results of the investigation as a whole, it may be concluded that the ignition of a solid explosive by a heated filament embedded in the medium is a purely thermal process. When suitable precautions are taken to avoid conversion of heat into work, the energy equation takes a simple form which may be interpreted as showing that the critical factor, in so far as the explosive is concerned, is the temperature. The geometry of the heating element probably determines the amount of explosive which must be raised to the critical temperature to ensure ignition. These criteria are independent of the time factor, and the increase in ignition energy observed with increasing time of heating is attributable to heat losses sustained by the ignition system during the heating period.

Strictly speaking, these conclusions apply only to ignition under the conditions specified, and it is well to bear in mind the limitations of the present investigation. In the first place, the experiments were confined to a range of solid explosives, and an extension of the work to other physical states of matter would seem desirable. Moreover, as the solids were bonded together to form a rigid matrix, the ignition system may be described as static, whereas ignition often occurs under conditions which, for contrast, might be termed dynamic. To quote the illustration used in the

introduction, heat cannot be generated in a gas at constant pressure without at the same time causing expansion and doing external work. In such a case the energy equation must contain the appropriate thermodynamic terms. Furthermore, the present model is based on a line source of heat, and it would facilitate mathematical generalization if data were available for another simple model, such as a point source or plane surface.

Granting that the above conclusions have only a limited application, it is nevertheless very remarkable that the energy equation for ignition appears to contain no term involving the heat of reaction. It would appear, therefore, that the heat of reaction does not enter into the process of ignition to any significant extent, and it must be concluded that, under these conditions of ignition, the specified criteria must be completely fulfilled before the exothermic phase of the reaction begins. Presumably, therefore, ignition occurs suddenly and simultaneously throughout the ignition system whenever the critical state is reached.

In conclusion, the author wishes to acknowledge that the experiments described in this paper were designed and carried out by Dr H. P. Stout, who was also largely responsible for the development of the experimental methods. Thanks are also due to Messrs G. Richards, I. G. Cumming and W. C. Warnock for assistance given at various stages in the investigation.

REFERENCE

Morgan, J. D. 1925 *Phil. Mag.* 49, 323.

Heisenberg's S matrix for a system of many particles

By R. J. EDEN, *Peterhouse, University of Cambridge*

(Communicated by P. A. M. Dirac, F.R.S.—Received 12 March 1949)

When creation and annihilation of particles is forbidden, the general features of a many-particle collision are illustrated by a scattering system containing two particles and a fixed scattering centre. The term in the wave function, which is called by Møller the 'outgoing' part, is shown to contain a term representing a totally outgoing wave, but also terms describing the interference effects between the incident wave of one particle and the outgoing wave of the other. Corresponding to the latter terms, there exist singular eigenstates of the S matrix which are simultaneous eigenstates of the separate kinetic energies of the particles. The remaining eigenstates are called non-singular, and for these only the sum of the kinetic energies can be given a definite value. Analytic continuation of the non-singular eigenstates in the complex plane of total kinetic energy shows that the corresponding eigenvalues of S can be used to determine the energy levels of states with both particles bound to the centre of force. The eigenvalues for the singular eigenstates will lead to the bound energy of a single particle in the scattering field of the other. The formalism is extended to include singular eigenstates which describe the scattering of one particle on a compound centre made up of the other particle bound to the scatterer.

1. INTRODUCTION

The usual form of quantum theory contains many quantities which are not observable. These quantities are introduced as auxiliary variables for convenience in calculation and are only indirectly related to experimental results. Examples are wave functions or dynamical variables which are written as functions of space-time and therefore can be exactly localized. Such localizability would be invalid in a physical theory if, for instance, there is some universal constant which plays the part of a minimal length.

Heisenberg (1949) has suggested that the divergence difficulties of quantum-field theories are a direct consequence of the use of localizable variables. However, it seems essential that those parts of quantum theory which are directly related to experimentally observed quantities must also be incorporated in any new theory of elementary particles. The theory of the characteristic matrix, or S matrix, is an attempt to set up a framework for a future theory which will contain none of the divergences but all the experimental results of present quantum mechanics.

The S matrix is based on the idea of a stationary collision in a system of particles, and its matrix elements are closely connected with collision cross-sections for various processes. The original formulation of the theory was given by Heisenberg (1943 *a, b*), and the general properties of S have been studied by Møller (1945, 1946). There are two main approaches to the development of S -matrix theory. One is to study the properties of the S matrix obtained from non-relativistic quantum mechanics, and show that it will lead to all experimental results which have previously been obtained from a Hamiltonian. In addition to cross-sections these include the energy levels of bound states of two or more particles and radio-active decay constants. The second approach is to obtain properties, from non-relativistic quantum mechanics, which are restrictions on the S matrix itself. Such properties must be independent of the

particular form of Hamiltonian, and may then be assumed to apply even if no Hamiltonian exists. The most important of these are the unitary condition which enables S to be interpreted as a transformation matrix, and the relativistic invariance of the eigenvalues of S . Both these are general conditions; the outstanding requirement is a special condition, which depends on the particular dynamical system under consideration. If S can be determined from these conditions, aided by a need for simplicity, they would form the basis for a new theory.

In this paper the author will be concerned principally with the first method of approach. The properties of a many-particle system are studied, with particular reference to the energy levels of bound states.

The Lorentz invariance properties of S have been studied by Møller (1945) for a collision of two particles, and he has made a formal extension to a many-particle collision. If one is not concerned with Lorentz invariance, the co-ordinates of the centre of gravity may be eliminated. The system is then mathematically equivalent to one in which particles are scattered on a fixed centre of force, with conservation of energy but not of momentum. Throughout this paper the author works in terms of a system of this type.

In their papers on S -matrix theory, Heisenberg (1943*a*) and Møller (1945) have assumed that the wave function for scattering may be put in a standard form having two parts, one representing the incident wave, and the other the 'outgoing' wave. For a system of one particle and a fixed scattering centre, Dirac (1948) has given the justification for this standard form of the wave function. The 'outgoing' wave does in fact correspond to the particle moving outward in the asymptotic region of co-ordinate space. It is shown in this paper, for a many-particle system, that the assumption of the standard form is related to the physical assumption that the particles interact only in a finite region of space. This implies that no Coulomb forces are present. The 'outgoing' part of the wave function corresponds to a state in which at least one particle is moving outward in the asymptotic region of co-ordinate space.

For the scattering of one particle, Heisenberg (1944) has shown that analytic continuation of the eigenvalues of S , in the complex energy plane, will lead to the energy levels of closed states in which the particle is bound to the scattering centre. The main purpose of the present paper is to extend the work of Heisenberg to systems of two or more particles with a scattering centre. It is assumed throughout that the number of particles is a constant of the motion. A system in which particles can be created and annihilated seems to present S -matrix theory with major difficulties which can probably be overcome only by new assumptions. Some study of this problem has been made by Hu (1949), but his new assumptions lead to a serious negative-energy difficulty.

It is shown first that the wave function, for scattering of two particles with no mutual interaction on a fixed centre, may be put into the standard form. The S matrix for the compound system is the product of the separate S matrices. This result can be extended to an arbitrary number of particles. After considering two particles having small mutual interaction, the interpretation of the 'outgoing' part of the wave function is examined. This is made up of interference terms between

'incident' waves of one particle and outgoing waves of the other, as well as the totally outgoing wave. The 'incident' wave of a particle is defined as the state in which this particle does not interact with the rest of the system; it will contain both incoming and outgoing parts. The product of the incident states of all the particles is called the initial state of the system.

In order to investigate bound states of a system containing two particles and a scattering centre, a superposition of states is formed. The basic states of the superposition correspond to incident plane waves of equal energies. The weight function for the superposition can be chosen so that the total state is an eigenstate of the S matrix. Then the weight function can be called an eigenfunction of S ; it determines the initial state of the system, which differs from the final state only by a phase factor. This phase factor is an eigenvalue of S .

Since the total energy W of the system commutes with S , every eigenfunction of S and W will contain a δ -function factor in the total energy of the two particles. Physically this means that, in a collision which is an eigenstate of S , one can know exactly the asymptotic value of the total kinetic energy. If, for a particular eigenstate, one knows the separate asymptotic values of the kinetic energies of the two particles, the eigenfunction will contain two δ -function factors, one for each kinetic energy. I call this a singular eigenfunction. It corresponds to a state in which the exchange of energy between the two particles is zero, so that asymptotically the separate kinetic energies are conserved. A necessary condition for the existence of singular eigenfunctions is that S contains terms having two δ -function factors in the separate kinetic energies. Such terms are always present in S , either as a product of incident waves, or as interference waves between one particle incident and the other outgoing. An eigenfunction which contains only the one δ -function factor in total kinetic energy is called a non-singular eigenfunction. The existence of such eigenfunctions is ensured if the separate kinetic energies do not commute with S .

By obtaining the asymptotic form of the wave function in co-ordinate space, we can investigate the conditions for bound states. For a non-singular eigenstate of S , analytic continuation of the corresponding eigenvalue, in the complex plane of total kinetic energy, leads to the energy levels of states with both particles bound to the centre of force. For a singular eigenstate, keeping the kinetic energy of one particle fixed, analytic continuation of the eigenvalue leads to the bound energy of the other particle.

The important case, where the fixed energy value is the bound energy of one particle, can be included in this scheme. It corresponds to a state in which one particle is scattered on a compound centre made up of the other particle bound to the scatterer. Since it is a singular eigenstate, ionization of the bound particle does not take place. When the total energy of the system is too low to allow ionization, the only possible states are a superposition of these singular eigenstates. With their corresponding eigenvalues these states define an S matrix which I call the 'partial' S matrix. It is not analytic and indicates new difficulties in the theory which are considered in more detail in another paper.

The possibility of a mixed eigenfunction, containing both singular and non-singular parts, is considered. It is shown that these parts may be regarded as separate

eigenfunctions of S , belonging to the same eigenvalue. This has the consequence that the singular part must correspond to a state in which one particle is bound.

Finally, some aspects of the extension to an arbitrary number of particles are discussed. It seems unlikely that any new points of principle would arise in making this extension.

2. TWO-PARTICLE SCATTERING WITH NO MUTUAL INTERACTION

Units are taken in which $\hbar = c = 1$. Denote the mass, momentum, and kinetic energy of a particle by $\kappa_n, \mathbf{k}_n, W_n$, for $n = 1, 2$. Then $W_n^2 = \kappa_n^2 + k_n^2$.

We consider two-particle scattering on a fixed centre of force, so that energy but not momentum must be conserved. Since there is no mutual interaction, the system can be treated as the product of two single-particle scattering systems. In the momentum space of one system the incident wave can be represented by a ket vector $|\mathbf{k}_n^A\rangle$, denoting a plane wave with momentum \mathbf{k}_n^A . Then the wave function takes the form

$$\langle \mathbf{k}_n | \psi_n | \mathbf{k}_n^A \rangle = \delta(\mathbf{k}_n - \mathbf{k}_n^A) + \delta_+(W_n - W_n^A) \langle \mathbf{k}_n | r_n | \mathbf{k}_n^A \rangle, \quad (2.1)$$

$$\text{where} \quad \delta_{\pm}(x) = \frac{\pm 1}{2\pi i x} + \frac{1}{2} \delta(x), \quad (2.2)$$

and $\delta(x)$ denotes Dirac's δ function.

For all possible initial values \mathbf{k}_n^A of the momentum, and all possible values \mathbf{k}_n , the wave functions (2.1) form the wave matrix ψ_n . Since the particles 1 and 2 have no mutual interaction, the wave function for the two-particle system is

$$\langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 | \psi_1 | \mathbf{k}_1^A \rangle \langle \mathbf{k}_2 | \psi_2 | \mathbf{k}_2^A \rangle, \quad (2.3)$$

where the initial state is $|\mathbf{k}_1^A\rangle |\mathbf{k}_2^A\rangle = |\mathbf{k}_1^A \mathbf{k}_2^A\rangle$.

Since, in forming (2.3), we shall be concerned with products of improper functions, it is necessary to use a more rigorous definition than (2.2) for $\delta_{\pm}(x)$. This is defined only when taken under an integral with respect to x , by

$$\int_{-\infty}^{\infty} \delta_{\pm}(a) da = \frac{1}{2\pi i} \lim_{\sigma \rightarrow +0} \int_{-\infty}^{\infty} \frac{da}{\pm a + i\sigma}, \quad (2.4)$$

$$\int_{-\infty}^{\infty} \delta(a) da = \int_{-\infty}^{\infty} \{\delta_+(a) + \delta_-(a)\} da = \frac{1}{2\pi i} \lim_{\sigma \rightarrow +0} \int \frac{2i\sigma da}{a^2 + \sigma^2}. \quad (2.5)$$

These are equivalent to the usual definitions. In a product of δ functions, the limit $\sigma \rightarrow 0$ must be taken last, after all integrations have been carried out. The following formulae are required:

$$\int \delta_+(a+b) [\delta_+(a) + \delta_+(b)] = \int \delta_+(a) \delta_+(b), \quad (2.6)$$

$$\int \delta(a+b) [\delta_+(a) + \delta_+(b)] = \int \delta(a) \delta(b), \quad (2.7)$$

where the \int sign denotes $\iint da db$ and a and b are independent variables. Using (2.4) the left side of (2.6) gives

$$\int \delta_+(a) \delta_+(b) - \frac{1}{4\pi^2} \lim_{\sigma \rightarrow +0} \int \frac{i\sigma}{(a+i\sigma)(b+i\sigma)(a+b+i\sigma)}.$$

The last term is zero, since the integrand is of order $1/a^2$ as $a \rightarrow \infty$ in any direction, and the integrand has no poles in the upper half-complex a plane for $\sigma > 0$. The left side of (2.7) gives

$$\int \delta(a+b) \delta(a) - \frac{1}{4\pi^2} \lim_{\sigma \rightarrow +0} \int \frac{2i\sigma(a+b)}{(a+b+i\sigma)(a+b-i\sigma)(a-i\sigma)(b+i\sigma)}.$$

The last term gives zero when the integration is carried out with respect to a and b . This proves (2.7). With the aid of (2.6) the wave function (2.3) can be written

$$\langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | 1 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle + \delta_+(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle, \quad (2.8)$$

where

$$\begin{aligned} \langle \mathbf{k}_1 \mathbf{k}_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle &= \delta(\mathbf{k}_1 - \mathbf{k}_1^A) \langle \mathbf{k}_2 | r_2 | \mathbf{k}_2^A \rangle + \delta(\mathbf{k}_1 - \mathbf{k}_2^A) \langle \mathbf{k}_1 | r_1 | \mathbf{k}_1^A \rangle \\ &\quad + [\delta_+(W_1 - W_1^A) + \delta_+(W_2 - W_2^A)] \langle \mathbf{k}_1 | r_1 | \mathbf{k}_1^A \rangle \langle \mathbf{k}_2 | r_2 | \mathbf{k}_2^A \rangle. \end{aligned} \quad (2.9)$$

In (2.8) $W = W_1 + W_2 = \sqrt{(k_1^2 + \kappa_1^2)} + \sqrt{(k_2^2 + \kappa_2^2)}$ is the total kinetic energy. The S matrix corresponding to (2.8) is defined by

$$\langle \mathbf{k}_1 \mathbf{k}_2 | S | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | 1 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle + \delta(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle \quad (2.10)$$

$$= \langle \mathbf{k}_1 | S_1 | \mathbf{k}_1^A \rangle \langle \mathbf{k}_2 | S_2 | \mathbf{k}_2^A \rangle. \quad (2.11)$$

In obtaining (2.11) we have used (2.7). In symbolic form this equation may be written

$$S = S_1 S_2. \quad (2.12)$$

This result is intuitively obvious, so that any other would cast doubt on the applicability of the theory to this particular system. But its proof is not trivial; for example, Heitler's theory of radiation damping leads to a different result, and must therefore be inapplicable to this problem. Heitler's integral equation, according to Wentzel (1947), can be written

$$R = 2\pi i K + \pi i K R, \quad (2.13)$$

where the matrix elements of K are the first non-vanishing terms of an interaction matrix, and

$$S = 1 + R. \quad (2.14)$$

In our example (2.13) holds for the product system, and since there is no interaction between the particles,

$$K = K_1 + K_2, \quad (2.15)$$

$$R_n = 2\pi i K_n + \pi i K_n R_n \quad (\text{for } n = 1, 2). \quad (2.16)$$

This gives

$$R = (R_1 + R_2 + R_1 R_2) / (1 - \frac{1}{2} R_1 R_2), \quad (2.17)$$

which contradicts (2.12).

The result (2.12) cannot be directly generalized to a system of many particles by induction, since the r matrix for many particles will contain a number of subsidiary terms with δ and δ_+ factors, which might invalidate our calculations. The direct proof will be outlined here. (2.6) and (2.7) generalize to

$$\int \delta_+ \left(\sum_{l=1}^n a_l \right) \sum_{m=1}^n \prod_{l \neq m}^n \delta_+(a_l) = \int \prod_{l=1}^n \delta_+(a_l), \quad (2.18)$$

$$\int \delta \left(\sum_{l=1}^n a_l \right) \sum_{m=1}^n \prod_{l \neq m}^n \delta_+(a_l) = \int \prod_{l=1}^n \delta(a_l), \quad (2.19)$$

where the \int sign denotes $\iint_{(n)} \int da_1 da_2 \dots da_n$, and a_1, a_2, \dots, a_n are independent. Left side of (2.18)

$$= \frac{1}{(2\pi i)^n} \left\{ \frac{1}{\prod_1^n (a_l + i\sigma)} + \frac{(n-1)i\sigma}{\left(\sum_1^n a_l + i\sigma\right) \prod_1^n (a_l + i\sigma)} \right\}.$$

The second term in the integrand is of order $1/a_l^2$ for any a_l , as $a_l \rightarrow \infty$ in the complex plane. It has no poles in the upper half a_l plane, and gives zero contribution to the integral. (2.18) follows. If (2.19) holds for n , then

$$\begin{aligned} & \int \delta(a_{n+1}) \delta\left(\sum_{l=1}^n a_l\right) \sum_{m=1}^n \prod_{l \neq m}^n \delta_+(a_l) = \int \prod_{l=1}^{n+1} \delta(a_l). \\ \text{Left side} &= \int \{\delta_+(a_{n+1}) + \delta_-(a_{n+1})\} \delta\left(\sum_{l=1}^{n+1} a_l\right) \sum_{m=1}^n \prod_{l \neq m}^n \delta_+(a_l) \\ &= \int \left\{ \delta_+(a_{n+1}) + \delta_+\left(\sum_{l=1}^n a_l\right) \right\} \delta\left(\sum_{l=1}^{n+1} a_l\right) \sum_{m=1}^n \prod_{l \neq m}^n \delta_+(a_l) \\ &= \text{left side of (2.19) with } n+1 \text{ for } n. \end{aligned}$$

It follows that (2.19) is true for all n .

The wave function for a system of n non-interacting particles, written in the product-momentum space, takes the form

$$\langle \mathbf{k}_1 \dots \mathbf{k}_n | \psi | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle = \prod_{l=1}^n \langle \mathbf{k}_l | \psi_l | \mathbf{k}_l^A \rangle. \quad (2.20)$$

Using (2.18) this can be written

$$= \langle \mathbf{k}_1 \dots \mathbf{k}_n | 1 | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle + \delta_+(W - W^A) \langle \mathbf{k}_1 \dots \mathbf{k}_n | r | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle. \quad (2.21)$$

The S matrix is given by

$$\langle \mathbf{k}_1 \dots \mathbf{k}_n | S | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle = \langle \mathbf{k}_1 \dots \mathbf{k}_n | 1 | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle + \delta(W - W^A) \langle \mathbf{k}_1 \dots \mathbf{k}_n | r | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle. \quad (2.22)$$

With the help of (2.19) this gives

$$S = \prod_{l=1}^n S_l. \quad (2.23)$$

3. TWO-PARTICLE SCATTERING WITH SMALL MUTUAL INTERACTION

When the interaction between two particles is treated as a small perturbation on the system of § 2, we can investigate the new form taken by the S matrix. Denote the unperturbed wave matrix by ψ_0 . Its matrix elements satisfy a wave equation

$$(W^A - W) \langle \mathbf{k}_1 \mathbf{k}_2 | \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | V \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle, \quad (3.1)$$

where $V = V_1 + V_2$ is the sum of the separate interactions of the particles with the scattering centre. Let the wave function $\langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle$ describe the state which develops adiabatically under a perturbing potential v . v refers to a small interaction between particles 1 and 2. Then

$$(W^A - W) \langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | (V + v) \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (3.2)$$

Put $\psi = \psi_0 + \phi$. Then to first order in v and ϕ ,

$$(W^A - W) \langle \mathbf{k}_1 \mathbf{k}_2 | \phi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | V \phi + v \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (3.3)$$

Without loss of generality we can take $V \phi = 0$, since this involves only a re-choosing of the unperturbed state ψ_0 . To see this, put $\phi = \phi_1 + \phi_2$ where $V \phi_2 = 0$. Then ϕ_1 satisfies (3.1) if ϕ_2 satisfies

$$(W^A - W) \langle \mathbf{k}_1 \mathbf{k}_2 | \phi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | v \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (3.4)$$

Hence ϕ_1 can be added to the unperturbed wave function ψ_0 .

It is assumed that particles 1 and 2 interact only in a finite region of space surrounding the scattering centre. Then the correction ϕ to the wave matrix will be of the form

$$\langle \mathbf{k}_1 \mathbf{k}_2 | \phi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = -2\pi i \delta_+(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | v \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (3.5)$$

Hence

$$\begin{aligned} \langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle \\ = \langle \mathbf{k}_1 \mathbf{k}_2 | 1 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle - 2\pi i \delta_+(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | V \psi_0 + v \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \end{aligned} \quad (3.6)$$

On going over to the S matrix this gives

$$\langle \mathbf{k}_1 \mathbf{k}_2 | S | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | S_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle - 2\pi i \delta(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | v \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (3.7)$$

This gives for the perturbed r matrix

$$\begin{aligned} \langle \mathbf{k}_1 \mathbf{k}_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle &= -2\pi i \langle \mathbf{k}_1 \mathbf{k}_2 | v | \mathbf{k}_1^A \mathbf{k}_2^A \rangle \\ &\quad - 2\pi i \int \langle \mathbf{k}_1 \mathbf{k}_2 | v | \mathbf{k}_1^A \mathbf{k}_2' \rangle d\mathbf{k}_2' \delta_+(W_2' - W_2^A) \langle \mathbf{k}_2' | r_2 | \mathbf{k}_2^A \rangle \\ &\quad - 2\pi i \int \langle \mathbf{k}_1 \mathbf{k}_2 | v | \mathbf{k}_1' \mathbf{k}_2^A \rangle d\mathbf{k}_1' \delta_+(W_1' - W_1^A) \langle \mathbf{k}_1' | r_1 | \mathbf{k}_1^A \rangle \\ &\quad - 2\pi i \iint \langle \mathbf{k}_1 \mathbf{k}_2 | v | \mathbf{k}_1' \mathbf{k}_2' \rangle d\mathbf{k}_1' d\mathbf{k}_2' \{ \delta_+(W_1' - W_1^A) + \delta_+(W_2' - W_2^A) \} \\ &\quad \times \langle \mathbf{k}_1' | r_1 | \mathbf{k}_1^A \rangle \langle \mathbf{k}_2' | r_2 | \mathbf{k}_2^A \rangle \\ &\quad + \langle \mathbf{k}_1 \mathbf{k}_2 | r_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle, \end{aligned} \quad (3.8)$$

where r_0 is the unperturbed wave matrix given by (2.9). We see from (3.8) that the perturbation causes an essential change in the r matrix, even when the energy shell condition $W = W^A$ is applied. If $V = V_1 + V_2$ is also small, of the same order as v , then

$$S_0 \doteq 1 - 2\pi i \bar{V}, \quad (3.9)$$

where

$$\langle \mathbf{k}_1 \mathbf{k}_2 | \bar{V} | \mathbf{k}_1^A \mathbf{k}_2^A \rangle \doteq \delta(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | V | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (3.10)$$

Then

$$S \doteq S_0(1 - 2\pi i \bar{v}) \doteq (1 - 2\pi i \bar{v}) S_0, \quad (3.11)$$

since the Poisson bracket $[\bar{V}, \bar{v}]$ is now of second order in smallness.

Excluding Coulomb fields, there would not be interaction between the two particles at infinite separation from the scattering centre, unless v could lead to

binding between the two particles. If $\mathbf{k}_{12}^{(m)}$ denotes the momentum of the compound particle in state (m) , v will have matrix elements of the type $\langle \mathbf{k}_{12}^{(m)} | v | \mathbf{k}_1^A \mathbf{k}_2^A \rangle$. Then

$$\begin{aligned} \langle \mathbf{k}_{12}^{(m)} | S | \mathbf{k}_1^A \mathbf{k}_2^A \rangle &= -2\pi i \delta(W - W^A) \langle \mathbf{k}_{12}^{(m)} | v \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle \\ &= -2\pi i \delta(W - W^A) \iint \langle \mathbf{k}_{12}^{(m)} | v | \mathbf{k}'_1 \mathbf{k}'_2 \rangle d\mathbf{k}'_1 d\mathbf{k}'_2 \langle \mathbf{k}'_1 \mathbf{k}'_2 | \psi_0 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \end{aligned} \quad (3.12)$$

However, the perturbation does not give the whole change in S , since the incident wave also will have to be modified. We will not investigate these types of transition in this paper. There is nothing essentially new in the generalization of this calculation to more than two particles.

4. THE PHYSICAL INTERPRETATION OF THE WAVE FUNCTION

For simplicity it is assumed that no matrix elements of type (3.12) occur. Then the matrix elements of ψ , which are important at large separation from the scattering centre, are

$$\langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \langle \mathbf{k}_1 \mathbf{k}_2 | 1 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle + \delta_+(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle, \quad (4.1)$$

$$\langle \mathbf{k}_1 n_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle = \delta_+(W - W^A) \langle \mathbf{k}_1 n_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (4.2)$$

The expression (4.1) denotes the pure scattering part of ψ , and describes a state in which neither particle is captured by the centre. (4.2) denotes a state in which particle 2 has been captured by the scattering centre, and the excess kinetic energy carried off by particle 1. Since we shall only be concerned with the behaviour at large distances from the centre, kinetic energy must be nearly conserved, so no states with both particles captured can occur.

The assumption that the particles interact only in a finite region round the scattering centre means that the incident waves can be plane waves, and the initial state of the system is represented by $\langle \mathbf{k}_1 \mathbf{k}_2 | 1 | \mathbf{k}_1^A \mathbf{k}_2^A \rangle$. This also implies that no Coulomb forces are present, only short-range forces act between the particles. When the wave function corresponding to the initial state is subtracted from (4.1), we are left with a part which should correspond to a state which has been affected by the scattering centre. That is, one or more of the particles must be outgoing in the asymptotic region of co-ordinate space. By transforming to co-ordinate space it will be shown that this requirement is satisfied by the last term of (4.1).

For one-particle scattering the corresponding requirement is that the asymptotic wave function should represent an incident wave plus a wave which is outgoing only. For two or more particles, in a steady scattering process, there are interference effects between the incident wave of one particle and the outgoing wave of another particle. For example, in (2.8) this interference is represented by the term

$$\delta_+(W - W^A) \delta(\mathbf{k}_2 - \mathbf{k}_2^A) \langle \mathbf{k}_1 | r_1 | \mathbf{k}_1^A \rangle.$$

This denotes a plane wave in the momentum space of particle 2, coupled with a wave outgoing from the scattering centre for particle 1. We note that the incident wave $\delta(\mathbf{k}_2 - \mathbf{k}_2^A)$ is incoming on one side of the scattering centre and outgoing on the other

side. Since the system is in a stationary state, the probability of any state is constant in time, so there is no contradiction in the fact that incoming waves of one particle mix with outgoing waves of the other even at infinity. Without such mixing of incident and outgoing waves we would not obtain a unitary S matrix.

If in a many-particle system there is mutual interaction between the particles, there will be an exchange of energy at the scattering centre. Without this exchange of energy a totally outgoing wave would contain a product of $\delta_+(W_i - W_i^A)$ factors, one for each outgoing particle. By formula (2.18) this reduces to the standard form with a $\delta_+(W - W^A)$ factor multiplied by a term singular in the separate W_i at W_i^A . With the exchange of energy the essential factor $\delta_+(W - W^A)$ remains, but the other will be modified so that it is no longer singular in the separate W_i . We will see that, in the asymptotic region of co-ordinate space, such a term still represents a totally outgoing wave.

Summarizing these remarks for a two-particle system: The second term in the wave function (4.1) will contain terms falling into two classes. First, the interference between particle 1 outgoing and particle 2 incident, given by

$$\delta_+(W - W^A) \delta(\mathbf{k}_2 - \mathbf{k}_2^A) \langle \mathbf{k}_1 | r_1 | \mathbf{k}_1^A \rangle. \quad (4.3)$$

Second, a term which we shall see is totally outgoing,

$$\delta_+(W - W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | f | \mathbf{k}_1^A \mathbf{k}_2^A \rangle, \quad (4.4)$$

where $\langle \mathbf{k}_1 \mathbf{k}_2 | f | \mathbf{k}_1^A \mathbf{k}_2^A \rangle$ contains no singular parts such as δ or δ_+ terms in the separate particle energies. The representation of (4.3) and (4.4) in co-ordinate space is obtained by means of the transformation function

$$\langle \mathbf{r}_1 \mathbf{r}_2 | \mathbf{k}_1 \mathbf{k}_2 \rangle = (2\pi)^{-3} \exp(i\mathbf{k}_1 \mathbf{r}_1) \exp(i\mathbf{k}_2 \mathbf{r}_2). \quad (4.5)$$

In \mathbf{r}_2 co-ordinate space (4.3) gives a plane wave $\exp(i\mathbf{k}_2^A \mathbf{r}_2)$ with momentum equal to the initial value \mathbf{k}_2^A . The \mathbf{r}_1 dependent part is identical with Møller's one-particle system (Møller 1946), and for large r_1 has the asymptotic form

$$\frac{\exp(ik_1^A r_1)}{ir_1 \sqrt{(2\pi)} k_1^A \sqrt{(\Delta_1^A)}} \langle W_1^A \omega_1^+ | r_1 | \mathbf{k}_1^A \rangle, \quad (4.6)$$

where $\Delta_1 = \frac{\partial(W_1 \omega_1)}{\partial(\mathbf{k}_1)}$, ω_1^+ are angle variables in the direction of \mathbf{r}_1 , and Δ_1^A is the value of Δ_1 at $W_1 = W_1^A$. (4.6) represents an outgoing wave only in \mathbf{r}_1 co-ordinate space.

The most convenient method of transforming (4.4) to co-ordinate space is by means of the Heisenberg-Møller δ functions. These are written $\delta_\pm(W, W^A)$ and $\delta(W, W^A)$ to distinguish them from the Dirac δ functions. Their properties are set out by Møller (1946), and only those formulae which are needed for specific discussion are given here.

$\delta_\pm(W, W^A)$ are defined when under an integral by

$$\int \delta_\pm(W, W^A) f(W) dW = \frac{\pm 1}{2\pi i} \int_{C_\pm(W^A)} \frac{f(W) dW}{(W - W^A)}. \quad (4.7)$$

$C_{\pm}(W^A)$ are contours in the complex W plane shown in figure 1.

$C_+(W^A)$ passes from the rest mass κ below the value W^A , and to infinity along the real axis.

$C_-(W^A)$ passes above W^A , and to infinity along the real axis.

When W^A is real and greater than κ these formulae are equivalent to (2.4).

$$\int \delta(W, W^A) f(W) dW = \frac{1}{2\pi i} \int_{C(W^A)} \frac{f(W) dW}{(W - W^A)} = f(W^A). \quad (4.8)$$

$C(W^A)$ is a closed contour running clockwise round W^A .

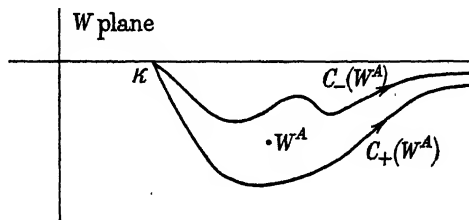


FIGURE 1

These δ functions are written as functions of two variables since the path of integration is modified to suit each value of W^A . The Dirac δ functions, on the other hand, are functions of only one variable given by the real difference $(W - W^A)$.

If $k^2 = W^2 - \kappa^2$ then for W^A real and greater than κ

$$\lim_{r \rightarrow \infty} \exp(ikr) \delta_{\pm}(W, W^A) = \begin{cases} \exp(ik^A r), \\ 0, \end{cases} \quad (4.9)$$

since $C_-(W^A)$ may be taken entirely in the upper half W plane, where the imaginary part of k is positive. Similarly, $C_+(W^A)$ may be taken in the upper half-plane except for a small closed contour surrounding the point $W = W^A$.

Transforming (4.4) to co-ordinate space we get for large r_1

$$\begin{aligned} & \frac{1}{(2\pi)^3} \iint d\mathbf{k}_1 d\mathbf{k}_2 \exp(i\mathbf{k}_1 \mathbf{r}_1 + i\mathbf{k}_2 \mathbf{r}_2) \delta_+(W, W^A) \langle \mathbf{k}_1 \mathbf{k}_2 | f | \mathbf{k}_1^A \mathbf{k}_2^A \rangle \\ &= \frac{1}{(2\pi)^2 i r_1} \iint \frac{\exp(ik_1 r_1)}{\sqrt{(\Delta_1) k_1}} \exp(i\mathbf{k}_2 \mathbf{r}_2) d\mathbf{k}_2 dW_1 \delta_+(W, W^A) \langle W_1 \omega_1^+ \mathbf{k}_2 | f | \mathbf{k}_1^A \mathbf{k}_2^A \rangle \\ & \quad - \frac{1}{(2\pi)^2 i r_1} \iint \frac{\exp(-ik_1 r_1)}{\sqrt{(\Delta_1) k_1}} \exp(i\mathbf{k}_2 \mathbf{r}_2) d\mathbf{k}_2 dW_1 \delta_+(W, W^A) \langle W_1 \omega_1^- \mathbf{k}_2 | f | \mathbf{k}_1^A \mathbf{k}_2^A \rangle, \end{aligned} \quad (4.10)$$

where ω_1^+ and ω_1^- are angle variables in the directions \mathbf{r}_1 and $-\mathbf{r}_1$ respectively. For any fixed value of \mathbf{k}_2 in the integrand of the last term, there is a factor

$$\exp(-ik_1 r_1) \delta_+(W_1, (W^A - W_2)).$$

The path of integration may be varied in the W_1 plane so that this gives zero, using the conjugate complex of equation (4.9). We now see why it is essential that f should contain no singular factor like $\delta(W_1^A - W_1)$, since with such a term only one point of the path of integration of W_1 is important and (4.9) would not apply. This would

give a non-zero contribution from the incoming wave $\exp(-ik_1 r_1)$. For a non-singular f , the asymptotic form of (4.4) for r_1 large is

$$\frac{1}{(2\pi)^2 i r_1} \int_{\kappa_1}^{W^A - \kappa_1} dW_2 \int d\omega_2 \frac{\exp(i k_1^B r_1) \exp(i k_2 r_2)}{\sqrt{(\Delta_1^B) k_1^B} \sqrt{(\Delta_2)}} \langle W_1^B \omega_1^+ W_2 \omega_2 | f | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (4.11)$$

The superscript B denotes that \mathbf{k}_1 and \mathbf{k}_2 are, in the integrand, taken on the kinetic energy shell, i.e.

$$\sqrt{(\kappa_1^2 + k_1^{B^2})} + \sqrt{(\kappa_2^2 + k_2^2)} = W_1^B + W_2 = W^A. \quad (4.12)$$

Similarly, if we had carried out the integration first with respect to \mathbf{k}_2 , we would have obtained, for r_2 large,

$$\frac{1}{(2\pi)^2 i r_2} \int_{\kappa_1}^{W^A - \kappa_1} dW_1 \int d\omega_1 \frac{\exp(i k_1 r_1) \exp(i k_2^B r_2)}{\sqrt{(\Delta_1)} \sqrt{(\Delta_2^B) k_2^B}} \langle W_1 \omega_1 W_2^B \omega_2^+ | f | \mathbf{k}_1^A \mathbf{k}_2^A \rangle, \quad (4.13)$$

where

$$W_2^B = W^A - W_1. \quad (4.14)$$

Hence it can be said that a non-singular term like f , in the r matrix, corresponds to an outgoing wave in the asymptotic regions of both \mathbf{r}_1 and \mathbf{r}_2 space, that is, it represents a totally outgoing wave.

For scattering of n particles on a fixed centre of force no essentially new concepts are involved. The general form of wave function assumed by Møller (1945) is

$$\begin{aligned} \langle \mathbf{k}_1 \dots \mathbf{k}_n | \psi | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle \\ = \langle \mathbf{k}_1 \dots \mathbf{k}_n | 1 | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle + \delta_+ (W - W^A) \langle \mathbf{k}_1 \dots \mathbf{k}_n | r | \mathbf{k}_1^A \dots \mathbf{k}_n^A \rangle. \end{aligned} \quad (4.15)$$

The last term corresponds to a state in which at least one particle is moving away from the scattering centre. It can be split into terms, some of which refer to a mixture of incident waves of some particles with outgoing waves of others.

The matrix element (4.2), which corresponds to capture of particle 2, can be written

$$\delta_+ \{ W_1, (W^A - W_2^{(n)}) \} \langle \mathbf{k}_1 n_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle. \quad (4.16)$$

This represents an outgoing wave in \mathbf{r}_1 co-ordinate space, with kinetic energy asymptotically equal to $W^A - W_2^{(n)}$, where $W_2^{(n)}$ is the bound energy of particle 2 in state (n) . In \mathbf{k}_2 momentum space this term is by definition the Fourier transform of a term like

$$1/r_2 \exp(-|k_2^{(n)}| r_2).$$

5. THE ASYMPTOTIC WAVE FUNCTION FOR AN EIGENSTATE OF S

In the previous section the discussion was restricted to initial states made up of products of plane incident waves. As a preliminary to the investigation of bound states, we now consider a general initial state for scattering of two particles. We take as basic states $|\mathbf{k}_1^A \mathbf{k}_2^A\rangle$, and superpose with a weight function $u(\mathbf{k}_1^A \mathbf{k}_2^A) \delta(W^A - W^0)$. This weight function ensures that the kinetic energy of the initial state is W^0 . The new wave function, which replaces (4.1), is

$$\psi_u(\mathbf{k}_1 \mathbf{k}_2) = \iint \langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle d\mathbf{k}_1^A d\mathbf{k}_2^A u(\mathbf{k}_1^A \mathbf{k}_2^A) \delta(W^A - W^0). \quad (5.1)$$

The suffix u on ψ_u indicates that we are considering a particular state determined by the function u . The $\delta(W^A - W^0)$ ensures that only values of u on the kinetic energy shell $W^A = W^0$ are important. The factor $\langle \mathbf{k}_1 \mathbf{k}_2 | \psi | \mathbf{k}_1^A \mathbf{k}_2^A \rangle$ in the integrand is given by (4.1). Writing

$$\delta(W - W^0) = \delta_-(W - W^0) + \delta_+(W - W^0),$$

(5.1) can be put in the form

$$\psi_u(\mathbf{k}_1 \mathbf{k}_2) = u(\mathbf{k}_1 \mathbf{k}_2) \delta_-(W - W^0) + v(\mathbf{k}_1 \mathbf{k}_2) \delta_+(W - W^0), \quad (5.2)$$

where

$$v(\mathbf{k}_1 \mathbf{k}_2) = u(\mathbf{k}_1 \mathbf{k}_2) + \iint \langle \mathbf{k}_1 \mathbf{k}_2 | r | \mathbf{k}_1^A \mathbf{k}_2^A \rangle d\mathbf{k}_1^A d\mathbf{k}_2^A u(\mathbf{k}_1^A \mathbf{k}_2^A) \delta(W^A - W^0). \quad (5.3)$$

Except in the asymptotic limit, $\psi_u(\mathbf{k}_1 \mathbf{k}_2)$, given by (5.2), is not on the energy shell $W = W^0$; but $u(\mathbf{k}_1 \mathbf{k}_2)$ can take arbitrary values for $W \neq W^0$ without affecting $\psi_u(\mathbf{k}_1 \mathbf{k}_2)$. The symmetry of form in (5.2) shows that $v(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0)$ will determine the superposition for a final state, in the same way as $u(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0)$ gives the initial state. Multiplying both sides of (5.3) by $\delta(W - W^0)$ gives

$$v(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0) = \iint \langle \mathbf{k}_1 \mathbf{k}_2 | S | \mathbf{k}_1^A \mathbf{k}_2^A \rangle d\mathbf{k}_1^A d\mathbf{k}_2^A u(\mathbf{k}_1^A \mathbf{k}_2^A) \delta(W^A - W^0). \quad (5.4)$$

In the special case where $u(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0)$ is an eigenfunction of S , it may be written as a transformation function,

$$u(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0) = \langle \mathbf{k}_1 \mathbf{k}_2 | W^0 \alpha^0 \rangle, \quad (5.5)$$

where W, α are a complete set of commuting observables, which commute also with S . Hence $S = S(W, \alpha)$ can be written as a function of them, and its eigenvalues written $\lambda(W^0 \alpha^0)$. Throughout this paper λ will denote an eigenvalue of S ; it is always a function of W^0 , the total kinetic energy of the system. Equation (5.4) gives for this special case

$$\begin{aligned} v(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0) &= \iint \langle \mathbf{k}_1 \mathbf{k}_2 | S | \mathbf{k}_1^A \mathbf{k}_2^A \rangle d\mathbf{k}_1^A d\mathbf{k}_2^A \langle \mathbf{k}_1^A \mathbf{k}_2^A | W^0 \alpha^0 \rangle \\ &= \lambda(W^0 \alpha^0) \langle \mathbf{k}_1 \mathbf{k}_2 | W^0 \alpha^0 \rangle \\ &= \lambda u(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0). \end{aligned} \quad (5.6)$$

λ is of modulus unity, so the initial and final states differ only by a phase factor.

In making a transformation of $\psi_u(\mathbf{k}_1 \mathbf{k}_2)$ to co-ordinate space, we again have to consider two possibilities. One is that $u(\mathbf{k}_1 \mathbf{k}_2)$ is 'non-singular', that is, it contains no δ factor in W_1 or W_2 . If (5.6) holds, then $v(\mathbf{k}_1 \mathbf{k}_2)$ is also non-singular. Secondly, $u(\mathbf{k}_1 \mathbf{k}_2)$ may contain a factor $\delta(W_1 - W_1^0)$; then it is called a singular function.

When $u(\mathbf{k}_1 \mathbf{k}_2) \delta(W - W^0)$ is a non-singular eigenfunction, the wave function in co-ordinate space is

$$\psi_u(\mathbf{r}_1 \mathbf{r}_2) = u(\mathbf{r}_1 \mathbf{r}_2) + v(\mathbf{r}_1 \mathbf{r}_2), \quad (5.7)$$

where the asymptotic forms of u and v for large r_1 are

$$u(\mathbf{r}_1 \mathbf{r}_2) \sim -\frac{1}{(2\pi)^2 i r_1} \int_{\kappa_1}^{W^0 - \kappa_1} dW_2 \int d\omega_2 \frac{\exp(-ik_1^A r_1)}{\sqrt{(\Delta_1^A) k_1^A}} \frac{\exp(ik_2^A r_2)}{\sqrt{(\Delta_2)}} u(W_1^A \omega_1^+ \mathbf{k}_2), \quad (5.8)$$

$$v(\mathbf{r}_1 \mathbf{r}_2) \sim -\frac{1}{(2\pi)^2 i r_1} \int_{\kappa_1}^{W^0 - \kappa_1} dW_2 \int d\omega_2 \frac{\exp(ik_1^A r_1)}{\sqrt{(\Delta_1^A) k_1^A}} \frac{\exp(ik_2^A r_2)}{\sqrt{(\Delta_2)}} v(W_1^A \omega_1^+ \mathbf{k}_2). \quad (5.9)$$

ω_1^+ and ω_1^- are the directions of \mathbf{r}_1 and $-\mathbf{r}_1$ respectively. The superscript A denotes that, in the integrand, k_1^A and W_1^A are treated as functions of W_2 according to the energy shell relation

$$W_1^A = \sqrt{(\kappa_1^2 + k_1^{A2})} = W^0 - W_2. \quad (5.10)$$

For large r_2 we obtain

$$u(\mathbf{r}_1, \mathbf{r}_2) \sim -\frac{1}{(2\pi)^2 i r_2} \int_{\kappa_1}^{W^0 - \kappa_2} dW_1 \int d\omega_1 \frac{\exp(i\mathbf{k}_1 \mathbf{r}_1)}{\sqrt{(\Delta_1)}} \frac{\exp(-ik_2^A r_2)}{\sqrt{(\Delta_2^A)} k_2^A} u(\mathbf{k}_1 W_2^A \omega_2^-), \quad (5.11)$$

$$v(\mathbf{r}_1, \mathbf{r}_2) \sim -\frac{1}{(2\pi)^2 i r_2} \int_{\kappa_1}^{W^0 - \kappa_2} dW_1 \int d\omega_1 \frac{\exp(i\mathbf{k}_1 \mathbf{r}_1)}{\sqrt{(\Delta_1)}} \frac{\exp(ik_2^A r_2)}{\sqrt{(\Delta_2^A)} k_2^A} v(\mathbf{k}_1 W_2^A \omega_2^+), \quad (5.12)$$

where $W_2^A = W^0 - W_1$. The physical interpretation of (5.9) to (5.12) shows that $u(\mathbf{k}_1, \mathbf{k}_2) \delta_-(W - W^0)$ represents an incoming wave in both \mathbf{r}_1 and \mathbf{r}_2 co-ordinate space, while $v(\mathbf{k}_1, \mathbf{k}_2) \delta_+(W - W^0)$ represents an outgoing wave.

When u is a singular function, the weight function for the superposition will take the form

$$u(\mathbf{k}_1, \mathbf{k}_2) \delta(W - W^0) = x(\mathbf{k}_1, \mathbf{k}_2) \delta(W_1 - W_1^0) \delta(W_2 - W_2^0), \quad (5.13)$$

where $x(\mathbf{k}_1, \mathbf{k}_2)$ is assumed to be a non-singular function. Physically a weight function of the type (5.13) means that the superposition for the initial state is made up of waves each having $W_1 = W_1^0$ and $W_2 = W_2^0$. If an eigenfunction of S takes the form (5.13) it is called a singular eigenfunction. Then the weight function for the final state will take the same form:

$$\begin{aligned} v(\mathbf{k}_1, \mathbf{k}_2) \delta(W - W^0) &= \int \int \langle \mathbf{k}_1, \mathbf{k}_2 | S | \mathbf{k}_1^A, \mathbf{k}_2^A \rangle d\mathbf{k}_1^A d\mathbf{k}_2^A x(\mathbf{k}_1^A, \mathbf{k}_2^A) \delta(W_1^A - W_1^0) \delta(W_2^A - W_2^0) \\ &= \lambda(W_1^0, W_2^0, \beta^0) x(\mathbf{k}_1, \mathbf{k}_2) \delta(W_1 - W_1^0) \delta(W_2 - W_2^0). \end{aligned} \quad (5.14)$$

In § 8 the conditions for the existence of both singular and non-singular eigenfunctions of S are discussed, and it is shown that they are normally satisfied. A singular eigenfunction corresponds physically to a state in which there is no interchange of energy between particles 1 and 2. For such a state one is able to measure the asymptotic values W_1^0 , W_2^0 of the separate kinetic energies. This distinguishes it from a non-singular eigenstate where only the total kinetic energy W^0 can be known.

When the singular function u is not an eigenfunction of S ,

$$v(\mathbf{k}_1, \mathbf{k}_2) \delta(W - W^0) = y(\mathbf{k}_1, \mathbf{k}_2) \delta(W_1 - W_1^0) \delta(W_2 - W_2^0) + z(\mathbf{k}_1, \mathbf{k}_2) \delta(W - W^0). \quad (5.15)$$

Substituting this in (5.2) gives

$$\begin{aligned} \psi_u(\mathbf{k}_1, \mathbf{k}_2) &= \delta(W_1 - W_1^0) x(\mathbf{k}_1, \mathbf{k}_2) \delta_-(W_2 - W_2^0) + \delta(W_1 - W_1^0) y(\mathbf{k}_1, \mathbf{k}_2) \delta_+(W_2 - W_2^0) \\ &\quad + z(\mathbf{k}_1, \mathbf{k}_2) \delta_+(W - W^0). \end{aligned} \quad (5.16)$$

In (5.16) the wave function has been split into δ_\pm terms with $(W_2 - W_2^0)$ as argument. This is consistent, but has removed the symmetry shown in (5.13) between particles 1 and 2. However, this has no effect on ψ_u itself. The wave function is split in this way, so that later the bound states of particle 2 can be investigated by analytic continuation of W_2^0 . The last term in (5.16) is of non-singular type, so it will represent

an outgoing wave in both \mathbf{r}_1 and \mathbf{r}_2 co-ordinate space. When $u(\mathbf{k}_1\mathbf{k}_2)\delta(W-W^0)$ gives a singular eigenstate of S , this last term is zero, and

$$\psi_u(\mathbf{k}_1\mathbf{k}_2) = \delta(W_1 - W_1^0)x(\mathbf{k}_1\mathbf{k}_2)\delta_-(W_2 - W_2^0) + \delta(W_1 - W_1^0)y(\mathbf{k}_1\mathbf{k}_2)\delta_+(W_2 - W_2^0). \quad (5.17)$$

Transforming (5.17) to co-ordinate space gives

$$\psi_u(\mathbf{r}_1\mathbf{r}_2) = x(\mathbf{r}_1\mathbf{r}_2) + y(\mathbf{r}_1\mathbf{r}_2). \quad (5.18)$$

For large r_2 ,

$$x(\mathbf{r}_1\mathbf{r}_2) \sim -\frac{1}{(2\pi)^2 i r_2} \int_{\kappa_1}^{W^0 - \kappa_2} dW_1 \int d\omega_1 \frac{\exp(i\mathbf{k}_1\mathbf{r}_1)}{\sqrt{(\Delta_1)}} \frac{\exp(-ik_2^A r_2)}{\sqrt{(\Delta_2^A)} k_2^A} \delta(W_1 - W_1^0) x(\mathbf{k}_1 W_2^A \omega_2^-) \quad (5.19)$$

$$\sim -\frac{1}{(2\pi)^2 i r_2} \int d\omega_1 \frac{\exp(i\mathbf{k}_1\mathbf{r}_1)}{\sqrt{(\Delta_1)}} \frac{\exp(-ik_2^0 r_2)}{\sqrt{(\Delta_2^0)} k_2^0} x(\mathbf{k}_1 W_2^0 \omega_2^-). \quad (5.20)$$

In (5.19) $W_1 + W_2^A = W^0$, and in (5.20) $W_1 + W_2^0 = W^0$, that is, $W_1 = W_1^0$, the initial and conserved value. (5.19) has been obtained by taking an integral over W_2 , with the path of integration in the lower half-plane except for a small circle round W_2^A .

Using the eigenvalue equation (5.14), we get also for large r_2

$$y(\mathbf{r}_1\mathbf{r}_2) \sim -\frac{1}{(2\pi)^2 i r_2} \int d\omega_1 \frac{\exp(i\mathbf{k}_1\mathbf{r}_1)}{\sqrt{(\Delta_1)}} \frac{\exp(ik_2^0 r_2)}{\sqrt{(\Delta_2^0)} k_2^0} \lambda(W_1^0 W_2^0 \beta^0) x(\mathbf{k}_1 W_2^0 \omega_2^+). \quad (5.21)$$

Equations (5.20) and (5.21) show that the first term on the right of (5.17) represents asymptotically an incoming wave in \mathbf{r}_2 space, and the last term an outgoing wave in \mathbf{r}_2 space. It is easy to see that the two parts have no such interpretation in \mathbf{r}_1 space. Transforming the first term on the right of (5.17) for large r_1 ,

$$x(\mathbf{r}_1\mathbf{r}_2) \sim \frac{1}{(2\pi)^2 i r_1} \iint d\mathbf{k}_2 dW_1 \frac{\exp(ik_1 r_1)}{\sqrt{(\Delta_1)} k_1} \exp(i\mathbf{k}_2\mathbf{r}_2) \delta(W_1 - W_1^0) x(W_1 \omega_1^+ \mathbf{k}_2) \delta_-(W_2, W_2^0) \\ - \frac{1}{(2\pi)^2 i r_1} \iint d\mathbf{k}_2 dW_1 \frac{\exp(-ik_1 r_1)}{\sqrt{(\Delta_1)} k_1} \exp(i\mathbf{k}_2\mathbf{r}_2) \delta(W_1 - W_1^0) x(W_1 \omega_1^- \mathbf{k}_2) \delta_-(W_2, W_2^0). \quad (5.22)$$

In (5.22) only the point $W_1 = W_1^0$, of the path of integration of W_1 , is important. It is this fact that has given a factor $\delta_-(W_2, W_2^0)$ instead of $\delta_-(W, W^0)$. Hence, unlike the non-singular case, we cannot make a variation of the path of integration to give a zero contribution from either term. Thus (5.22) represents both incoming and outgoing waves in \mathbf{r}_1 co-ordinate space.

6. ANALYTIC CONTINUATION FOR A NON-SINGULAR EIGENSTATE

In this section the behaviour of the wave function (5.7) is investigated when analytic continuation is made in the complex W^0 plane, where W^0 is the total kinetic energy of the system. For a pure scattering problem, the asymptotic value of the kinetic energy W_i of each particle must be greater than its rest mass κ_i . Hence for scattering

$$\kappa_1 + \kappa_2 < W^0. \quad (6.1)$$

It is supposed that the wave function (5.7) is known for the range (6.1), and the effect on the asymptotic forms (5.8) to (5.12) is investigated when W^0 is continued analytically to the range

$$0 < W^0 < \kappa_1 + \kappa_2. \quad (6.2)$$

In the equations (5.8) to (5.12) both the integrand and also the path of integration are functions of W^0 . This enables us to obtain the behaviour of the asymptotic wave function when W^0 becomes complex, provided we specify along what path W^0 is continued. This will define the new path of integration for W_1 or W_2 . W_1 and W_2 satisfy

$$W_1 = \sqrt{(\kappa_1^2 + k_1^2)}, \quad W_2 = \sqrt{(\kappa_2^2 + k_2^2)}, \quad (6.3)$$

$$W_1 + W_2 = W_0. \quad (6.4)$$

The W_i plane must be cut from κ_i to $-\kappa_i$ in order to make W_i a single-valued function of the momentum k_i . We adopt the convention that k_1 and k_2 are both real and positive for W^0 in range (6.1). This fixes the Riemann sheet in which we start.

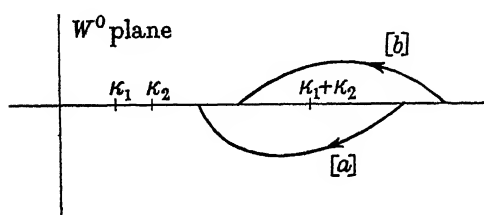


FIGURE 2

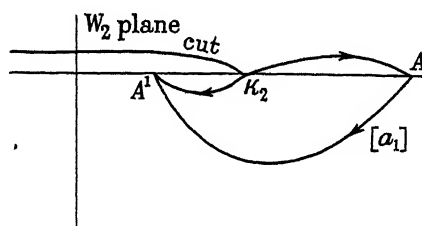


FIGURE 3

We consider the behaviour of (5.8) when W^0 is continued by path [a] (figure 2). to values in range (6.2). The path of integration in the W_2 plane depends linearly on this change in W^0 . It changes from κ_2 to A , by the addition of a path [a₁] (figure 3). If it is assumed that the integrand of (5.8) has no singularities in the lower half W_2 plane, the path of integration may be contracted to the direct path $\kappa_2 A^1$, taken *below* the cut on the real axis (κ_2 to $-\kappa_2$). Since $k_2 = |W_2^2 - \kappa_2^2|^{\frac{1}{2}} \exp(\frac{1}{2}i\theta)$ and $0 > \theta > -\pi$ on path [a₁], we have $k_2 = |k_2| \exp(-\frac{1}{2}i\pi)$ along path $\kappa_2 A^1$. For fixed W^0 given by the value $\kappa_1 + A^1$, as W_2 varies along the path $\kappa_2 A^1$, W_1^A in the integrand of (5.8) will be given by $W_1^A = W^0 - W_2$. W^0 is real and W_2 is in the lower half-complex plane. Hence in the integrand of (5.8), $k_1^A = +i|k_1^A|$ as W_2 goes along $\kappa_2 A^1$. This means that for large r_1 , the wave function $u(r_1 r_2)$ given by (5.8) contains a diverging factor $\exp(-ik_1^A r_1) = \exp(+|k_1^A| r_1)$ when W^0 is continued by path [a] to real values in range (6.2). Similarly $v(r_1 r_2)$ in (5.9) contains a convergent factor $\exp(-|k_1^A| r_1)$ for the same continuation of W^0 . For large r_2 , we have to consider (5.11) and (5.12). In these W_1 is the free variable whose path of integration κ_1 to $W^0 - \kappa_2$ is affected directly by continuation of W^0 . When W^0 is continued by path [a] (figure 2), W_1 has a path of integration from κ_1 and below the cut (κ_1 to $-\kappa_1$) to $W^0 - \kappa_2 < \kappa_1$. For this path the dependent variable $W_2^A = W^0 - W_1$ moves in the upper half-plane, hence $k_2^A = +i|k_2^A|$, in the integrands of (5.11) and (5.12).

Summarizing these results for continuation of W^0 by path [a] (figure 2),

$$\left. \begin{aligned} u(r_1 r_2) &\sim r_1^{-1} \exp(|k_1^A| r_1) && \text{for large } r_1, \\ &r_2^{-1} \exp(|k_2^A| r_2) && \text{for large } r_2, \end{aligned} \right\} \quad (6.5)$$

$$\left. \begin{aligned} v(r_1 r_2) &\sim r_1^{-1} \exp(-|k_1^A| r_1) && \text{for large } r_1, \\ &r_2^{-1} \exp(-|k_2^A| r_2) && \text{for large } r_2, \end{aligned} \right\} \quad (6.6)$$

We see that $v(\mathbf{r}_1 \mathbf{r}_2)$ represents a closed state, in which particles 1 and 2 are bound to the scattering centre. If continuation of W^0 is made by path [b] (figure 2) above the real axis then

$$\begin{aligned} u(\mathbf{r}_1 \mathbf{r}_2) &\sim r_1^{-1} \exp(-|k_1^A| r_1) \quad \text{for large } r_1, \\ &\sim r_2^{-1} \exp(-|k_2^A| r_2) \quad \text{for large } r_2. \end{aligned} \quad (6.7)$$

$$\begin{aligned} v(\mathbf{r}_1 \mathbf{r}_2) &\sim r_1^{-1} \exp(|k_1^A| r_1) \quad \text{for large } r_1, \\ &\sim r_2^{-1} \exp(|k_2^A| r_2) \quad \text{for large } r_2. \end{aligned} \quad (6.8)$$

In this case, $u(\mathbf{r}_1 \mathbf{r}_2)$ represents a closed state with particles 1 and 2 bound.

For continuation by path [a], if the wave function $\psi_u(\mathbf{r}_1 \mathbf{r}_2)$ is itself to represent a closed state, we must have $u(\mathbf{r}_1 \mathbf{r}_2)$ zero. This will happen when there is a zero term in the integrands of (5.8) and (5.11), that is

$$u(W_1^A \omega_1^- \mathbf{k}_2) \equiv 0 \quad \text{and} \quad u(\mathbf{k}_1 W_2^A \omega_2^-) \equiv 0. \quad (6.9)$$

These functions must be zero for all values of the variables, subject to $W_1^A + W_2 = W^0$, and $W_1 + W_2^A = W^0$ respectively. Hence the zero must be associated with the particular fixed final value, $W^{(n)}$ say, of W^0 , where $0 < W^{(n)} < \kappa_1 + \kappa_2$. From (5.6) we see that $W^{(n)}$ must be a root of

$$\lambda(W^0 \alpha^0) = \infty. \quad (6.10)$$

Similarly, if W^0 is continued by path [b] (figure 2) to the range (6.2), $W^{(n)}$ is found to be a root of

$$\lambda(W^0 \alpha^0) = 0. \quad (6.11)$$

$W^{(n)}$ is interpreted as the energy of a state in which both particles are bound to the scattering centre.

These results show that, providing there exist eigenfunctions of the S matrix which are non-singular in the kinetic energy of one particle alone, the energies of the closed states may be obtained by analytic continuation of the corresponding eigenvalues. It will be seen later that in general non-singular eigenfunctions do exist.

It is assumed that a correct S matrix, calculated by some future theory, will lead only to zeros which correspond to bound states, although it has been noted that redundant zeros with no physical meaning may occur when the S matrix is calculated from non-relativistic quantum mechanics.

7. ANALYTIC CONTINUATION FOR A SINGULAR EIGENSTATE

We now consider the effect on the wave function (5.18) when analytic continuation is made in the W_2^0 plane. In (5.22) we have seen that the term $x(\mathbf{r}_1 \mathbf{r}_2)$ contains both incoming and outgoing waves in \mathbf{r}_1 space, when the singular value is real and greater than κ_1 . The same is true for $v(\mathbf{r}_1 \mathbf{r}_2)$. We consider first $W_1^0 > \kappa_1$, and then $W_1^0 < \kappa_1$.

For large r_2 , analytic continuation of W^0 by path [a] (figure 2), corresponds to continuation of W_2^0 by path [a₂] (figure 4). This leads to $k_2^0 = i|k_2^0|$ in (5.20) and (5.21) giving

$$x(\mathbf{r}_1 \mathbf{r}_2) \sim r_2^{-1} \exp(|k_2^0| r_2), \quad (7.1)$$

$$y(\mathbf{r}_1 \mathbf{r}_2) \sim r_2^{-1} \exp(-|k_2^0| r_2). \quad (7.2)$$

Thus $x(\mathbf{r}_1 \mathbf{r}_2)$ diverges, whilst $y(\mathbf{r}_1 \mathbf{r}_2)$ represents a bound state of particle 2. We can see directly that when W_2^0 moves into the upper half-plane from B , the path of integration $C_+(W_2^0)$ can be taken entirely in the upper half-plane when associated with $\exp(ik_2^0 r_2)$, and entirely in the lower half-plane when associated with $\exp(-ik_2^0 r_2)$. Thus a $\delta_+(W_2, W_2^0)$ factor will lead to a bound state of particle 2, using path $[a_2]$. But $C_-(W_2^0)$ for W_2^0 at B^1 must encircle B^1 before going to infinity, so a $\delta_-(W_2, W_2^0)$ factor will in general lead to a divergent term.

If, for a particular final value $W_2^{(n)}$ of W_2^0 ,

$$x(\mathbf{k}_1 W_2^0 \omega_2) = 0 \quad \text{for} \quad k_2^0 = i |k_2^{(n)}|. \quad (7.3)$$

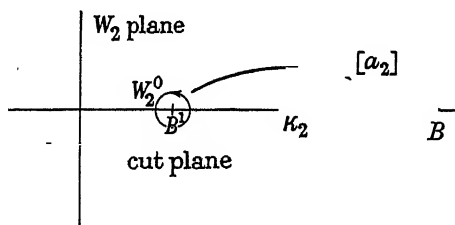


FIGURE 4

Then the wave function $\psi_u(\mathbf{r}_1 \mathbf{r}_2)$ will itself represent a state in which particle 2 is bound and particle 1 free. The eigenvalue $\lambda(W_1^0 W_2^0 \beta^0)$ which corresponds to a singular eigenstate is a function of both W_1^0 and W_2^0 , since for this state both kinetic energies can be known. Hence for each fixed value of W_1^0 , the bound states of particle 2 have energies $W_2^{(n)}$ satisfying

$$\lambda(W_1^0 W_2^{(n)} \beta^0) = \infty, \quad k_2^{(n)} = i |k_2^{(n)}|, \quad (7.4)$$

$$\lambda(W_1^0 W_2^{(n)} \beta^0) = 0, \quad k_2^{(n)} = -i |k_2^{(n)}|. \quad (7.5)$$

The solutions $W_2^{(n)}$ will in general be dependent on the particular fixed value of W_1^0 ; that is, $W_2^{(n)}$ is the energy of particle 2 bound to the centre and subject to the field of the scattered particle 1 of energy W_1^0 . For $W_1^0 > \kappa_1$ the dependence of $W_2^{(n)}$ on W_1^0 will probably be slight, but it is certainly important for $W_1^0 < \kappa_1$. The eigenvalues β^0 which refer to angle variables will also affect $W_2^{(n)}$.

When $W_1^0 = W_1^{(n)} < \kappa_1$ is the bound energy of particle 1 and the scatterer, the singular eigenstate has a special physical significance. It refers to scattering of particle 2 on a compound system made up of particle 1 bound to the scatterer. For the total energy $W^0 < \kappa_1 + \kappa_2$ all eigenstates are singular and of this type. A singular eigenstate of this special type can be written as a ket vector $|W_1^{(n)} W_2^0 \beta^0\rangle$. There are two ways of representing this state in momentum space. One is to take the usual Fourier transform from the \mathbf{r}_1 co-ordinate representation, but this has the disadvantage that $\langle \mathbf{k}_1 \mathbf{k}_2 | W_1^{(n)} W_2^0 \beta^0 \rangle$ does not contain a δ factor in the kinetic energy. The other is to use a complex momentum space denoted by $\langle * \mathbf{k}_1 \mathbf{k}_2 |$ in which W_1^0 may be less than κ_1 . Using this the eigenfunction takes the form

$$\langle * \mathbf{k}_1 \mathbf{k}_2 | W_1^{(n)} W_2^0 \beta^0 \rangle = \delta(W_1, W_1^{(n)}) \delta(W_2 - W_2^0) x(\mathbf{k}_1 \mathbf{k}_2). \quad (7.6)$$

The use of complex momentum space makes the treatment of singular eigenstates for $W_1^0 < \kappa_1$ formally the same as for $W_1^0 > \kappa_1$. The eigenfunction equation becomes

$$\iint \langle * \mathbf{k}_1 \mathbf{k}_2 | S | \mathbf{k}_1^A \mathbf{k}_2^A * \rangle d\mathbf{k}_1^A d\mathbf{k}_2^A \langle * \mathbf{k}_1^A \mathbf{k}_2^A | W_1^{(n)} W_2^0 \beta^0 \rangle \\ = \lambda(W_1^{(n)} W_2^0 \beta^0) \langle * \mathbf{k}_1 \mathbf{k}_2 | W_1^{(n)} W_2^0 \beta^0 \rangle. \quad (7.7)$$

Since we are considering a real scattering problem here (and not the analytic continuation of one) the eigenvalue λ is of modulus unity for $W_2^0 > \kappa_2$. If $\lambda(W_1^{(n)} W_2^0 \beta^0)$ is continued analytically to the region $0 < W_2^0 < \kappa_2$, the energies $W_2^{(m)}$ of the bound states of particle 2 are obtained from the solutions $W_2^0 = W_2^{(m)}$ of

$$\lambda(W_1^{(n)} W_2^0 \beta^0) = \infty, \quad k_2^0 = i |k_2^0|, \quad (7.8)$$

$$\lambda(W_1^{(n)} W_2^0 \beta^0) = 0, \quad k_2^0 = -i |k_2^0|. \quad (7.9)$$

The value $W_2^{(m)}$ is the energy of particle 2 bound, subject to particle 1 having energy $W_1^{(n)}$. Hence $W_1^{(n)} + W_2^{(m)} = W^{(n)}$ is the energy of a state with both particles bound to the scatterer.

8. FURTHER INVESTIGATION OF EIGENSTATES

We have the following commutation properties in general,

$$[S, W] = 0, \quad [S, W_1] = -[S, W_2] \neq 0. \quad (8.1)$$

All the eigenstates considered are simultaneous eigenstates of S and W . The singular eigenstates are simultaneous eigenstates of S , W_1 and W_2 . Not every eigenstate is a singular eigenstate unless there is zero interaction between the particles, except in certain bounded energy ranges.

The eigenvalue equation for a singular eigenfunction is

$$\iint \langle W_1 \omega_1 W_2 \omega_2 | S | W_1^0 \omega_1' W_2^0 \omega_2' \rangle d\omega_1' d\omega_2' x(W_1^0 \omega_1' W_2^0 \omega_2') \\ = \lambda(W_1^0 W_2^0 \beta^0) \delta(W_1 - W_1^0) \delta(W_2 - W_2^0) x(W_1 \omega_1 W_2 \omega_2). \quad (8.2)$$

Apply $\int_{W_1^0-\epsilon}^{W_1^0+\epsilon} dW_1 \int_{W_2^0-\epsilon}^{W_2^0+\epsilon} dW_2$ to both sides of (8.2) and let $\epsilon \rightarrow 0$. The right-hand side gives $\lambda(W_1^0 W_2^0 \beta^0) x(W_1^0 \omega_1 W_2^0 \omega_2)$, and the left-hand side gives zero unless it contains two δ function factors of the type $\delta(W_1 - W_1^0) \delta(W_2 - W_2^0)$. These always occur in $\langle W_1 \omega_1 W_2 \omega_2 | S | W_1^0 \omega_1' W_2^0 \omega_2' \rangle$, and arise both in the term corresponding to a product of incident waves, and in the term corresponding to interference between the incident wave of one particle and the outgoing wave of the other.

If there exist eigenstates which are a mixture of the singular and the non-singular types, they can be written

$$\delta(W - W^0) u(\mathbf{k}_1 \mathbf{k}_2) + \delta(W_1 - W_1^0) \delta(W_2 - W_2^0) x(\mathbf{k}_1 \mathbf{k}_2). \quad (8.3)$$

The eigenvalue $\lambda(W^0)$ corresponding to (8.3) is a particular function of W^0 having singular points at $W^0 = W^{(n)} < \kappa_1 + \kappa_2$. We cannot at once interpret $W^{(n)}$ as the bound

energy of two particles, since this result was obtained when $\lambda(W^0)$ was an eigenvalue of a non-singular eigenfunction. The eigenvalue equation is

$$\begin{aligned} \iint \langle \mathbf{k}_1 \mathbf{k}_2 | S | \mathbf{k}'_1 \mathbf{k}'_2 \rangle d\mathbf{k}'_1 d\mathbf{k}'_2 \{ \delta(W' - W^0) u(\mathbf{k}'_1 \mathbf{k}'_2) + \delta(W'_1 - W^0_1) \delta(W'_2 - W^0_2) x(\mathbf{k}'_1 \mathbf{k}'_2) \} \\ = \lambda(W^0) \{ \delta(W - W^0) u(\mathbf{k}_1 \mathbf{k}_2) + \delta(W_1 - W^0_1) \delta(W_2 - W^0_2) x(\mathbf{k}_1 \mathbf{k}_2) \}. \end{aligned} \quad (8.4)$$

If $W^0_1 > \kappa_1$ the range of integration of W'_1 will be along the real axis from κ_1 to $W^0 - \kappa_2$. But if $W^0_1 < \kappa_1$ the $\delta(W'_1 - W^0_1)$ factor has to be replaced by $\delta(W'_1, W^0_1)$ which requires the path to be a small circle round W^0_1 . Apply $\int_{W^0_1 - \epsilon}^{W^0_1 + \epsilon} dW_1 \int_{W^0_2 - \epsilon}^{W^0_2 + \epsilon} dW_2$ to equation (8.4). If $W^0_1 < \kappa_1$ we would apply an integral with respect to W_1 round the point W^0_1 and radius ϵ . Let $\epsilon \rightarrow 0$. The right-hand side gives $\lambda(W^0) x(\mathbf{k}_1 \mathbf{k}_2)$ with $W_1 = W^0_1$, $W_2 = W^0_2$. The first term on the left gives zero since it is not singular in both W_1 and W_2 . Hence

$$\begin{aligned} \int_{W^0_1 - \epsilon}^{W^0_1 + \epsilon} dW_1 \int dW_2 \iint \langle \mathbf{k}_1 \mathbf{k}_2 | S | \mathbf{k}'_1 \mathbf{k}'_2 \rangle d\mathbf{k}'_1 d\mathbf{k}'_2 \delta(W'_1 - W^0_1) \delta(W'_2 - W^0_2) x(\mathbf{k}'_1 \mathbf{k}'_2) \\ = \lambda(W^0) x(\mathbf{k}_1 \mathbf{k}_2), \end{aligned} \quad (8.5)$$

with $W_1 = W^0_1$, $W_2 = W^0_2$, in $x(\mathbf{k}_1 \mathbf{k}_2)$. This means we can take

$$\delta(W_1 - W^0_1) \delta(W_2 - W^0_2) x(\mathbf{k}_1 \mathbf{k}_2)$$

itself to be an eigenfunction of S , belong to the eigenvalue $\lambda(W^0)$. Hence

$$\delta(W - W^0) u(\mathbf{k}_1 \mathbf{k}_2)$$

is also an eigenfunction belonging to $\lambda(W^0)$. We can deduce that the singularities $W^{(n)}$ are the energies of bound states of the two particles. This has the consequence that $\delta(W_1 - W^0_1) \delta(W_2 - W^0_2) x(\mathbf{k}_1 \mathbf{k}_2)$ is a special singular eigenstate with either $W^0_1 = W^{(n)}_1 < \kappa_1$, or $W^0_2 = W^{(n)}_2 < \kappa_2$. Thus any eigenstate of a mixed type can always be represented as a superposition of a singular and a non-singular eigenstate. The singular state will in this case correspond to one of the particles bound to the scattering centre.

When the total energy of the system $W^0 < \kappa_1 + \kappa_2$, one particle must be bound to the scattering centre. This means that in any eigenstate the energy W_2 is conserved asymptotically, so the only possible eigenstates are of the type (7.7). These eigenstates vary in number according to the number of levels $W^{(n)}_1$ allowed by energy conservation, which requires $W^{(n)}_1 + \kappa_2 < W^0$. For each W^0 these singular eigenstates and their corresponding eigenvalues define an S matrix, which I call the 'partial S matrix' since it contains only a part of all possible scattering matrix elements. The concept of a partial S matrix is only relative, since for a general system in which particles can be created and annihilated, the size of the S matrix increases indefinitely with increasing energy. The partial S matrix is unitary in order to conserve probability in a collision, but its 'piece-wise' analytic behaviour gives rise to difficulties in the theory, which will be discussed in more detail in another paper.

9. CONCLUSION

The extension of the foregoing work to a system composed of an arbitrary constant number of particles and a fixed scattering centre should be possible without introducing any more new concepts. However, there are one or two points worth mentioning.

The non-singular eigenstates will refer to pure scattering of all the particles with mutual interchanges of energy which prevent any particle, or set of particles less than the complete system, from remaining on the kinetic energy shell. The wave function in co-ordinate space will be

$$\psi_u(\mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_n) = u(\mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_n) + v(\mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_n), \quad (9.1)$$

corresponding to (5.7). For asymptotically large r_1 ,

$$u(\mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_n) \sim \frac{-1}{(2\pi)^{-1+\frac{1}{2}n} i r_1} \int \dots \int dW_2 d\omega_2 \dots dW_n d\omega_n \frac{\exp(-ik_1^A r_1) \exp(i\mathbf{k}_2 \mathbf{r}_2)}{\sqrt{(\Delta_1^A)} k_1^A} \frac{\exp(i\mathbf{k}_n \mathbf{r}_n)}{\sqrt{(\Delta_n)}} u(W_1^A \omega_1 W_2 \omega_2 \dots W_n \omega_n). \quad (9.2)$$

In the integrand of (9.2), $W_1^A = W^0 - \sum_{i=2}^n W_i$. (9.3)

The path of integration of W_i for $i = 2, 3, \dots, n$, is κ_i to $W^0 - \sum_{j \neq i} \kappa_j$. Analytic continuation of W^0 by path [a] (figure 2) to a point just below the real axis, will cause the paths of integration of all W_i ($i = 2, 3, \dots, n$), to move down below the real axis. Hence W_1^A , given by (9.3) with W^0 equal to its final real value less than $\sum_1^n \kappa_i$, will move along a path having $k_1^A = i |k_1^A|$ at all points. Thus $u(\mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_n)$ diverges in r_1 , and similarly $v(\mathbf{r}_1 \mathbf{r}_2 \dots \mathbf{r}_n)$ converges in r_1 . The remaining work follows as before.

There will be many singular eigenfunctions, corresponding to kinetic energy conservation of any set of m particles where $m < n$. These eigenfunctions arise from the very complicated mixing between incident waves of some particles and outgoing waves of others, and also from the states in which some of the particles are bound to the scattering centre. The concept of a partial S matrix will still be valid, but it will be discontinuous at a large number of energy values.

I would like to express my thanks to Professor P. A. M. Dirac for valuable advice and criticism, and to Professor W. Heisenberg for many helpful discussions.

REFERENCES

- Dirac, P. A. M. 1948 *Principles of quantum mechanics*, 3rd ed. Oxford University Press.
 Heisenberg, W. 1943a Paper I. *Z. Phys.* **120**, 513.
 Heisenberg, W. 1943b Paper II. *Z. Phys.* **120**, 673.
 Heisenberg, W. 1944 Paper III. *Z. Phys.* **123**, 93.
 Heisenberg, W. Paper IV. Unpublished. (See Heisenberg 1946.)
 Heisenberg, W. 1946 *Z. Naturforsch.* **1**, 608.
 Heisenberg, W. 1949 *Two lectures*. Cambridge University Press.
 Hu, N. 1948 *Proc. R. Irish Acad.* **52**, A 5.
 Møller, C. 1945 *K. danske vidensk. Selsk. Math-Fys. Medd.* **23**, no. 1.
 Møller, C. 1946 *K. danske vidensk. Selsk. Math-Fys. Medd.* **22**, no. 19.
 Wentzel, G. 1947 *Rev. Mod. Phys.* **19**, 1.

Penetration of magnetic field into superconductors.

II. Measurements by the Casimir method

BY E. LAURMANN AND D. SHOENBERG

Royal Society Mond Laboratory, Cambridge

(Communicated by Sir Lawrence Bragg, F.R.S.—Received 14 March 1949—

Read 16 June 1949)

The changes with temperature of penetration of a magnetic field into superconducting tin and mercury were studied by a method due to Casimir in which a mutual inductance with a superconducting core is measured using low-frequency currents. The results were found to be very sensitive to surface conditions, but single crystals with smooth surfaces gave reproducible measurements of $\lambda(T) - \lambda(2.17^\circ \text{K})$ as a function of temperature T . These were consistent with the formula $\lambda(T) = \lambda_0(1 - (T/T_c)^4)^{-\frac{1}{2}}$, where T_c is the transition temperature, and λ_0 was found to be 5.2×10^{-6} cm. for tin and 4.3×10^{-6} cm. for mercury. For tin there was no significant difference between the values of λ_0 for current flow in different crystal directions, though a difference of up to 20 % is not excluded. For mercury there is a suggestion that λ_0 is about 20 % higher for current flow perpendicular to the principal axis than it is for current flow parallel to the principal axis, but this difference is little more than might be due to experimental errors. There was no evidence for any dependence of λ on a steady magnetic field H , though an increase of 10 % up to 80 % of the critical field is not excluded.

INTRODUCTION

The first attempt to study the depth of penetration of a magnetic field into a superconductor of macroscopic dimensions was made by Casimir (1940). The mutual inductance of two coils wound on a superconducting mercury core should become slightly smaller as the temperature is lowered, if the penetration depth decreases with temperature as suggested by earlier experiments on colloidal mercury (Shoenberg 1940). The expected decrease of mutual inductance was very small, of order $0.3 \mu\text{H}$ in the apparatus of Casimir, but with the sensitive bridge used should have been easily detected. In fact Casimir found no decrease that could be attributed to this effect and concluded that either there was a difference in the behaviour of macroscopic and colloidal specimens, or the superconductor did not behave in the same way at the frequency of measurement (about 80 c./sec.) as in the static conditions of the colloid experiments.

Since this question is of considerable importance for the phenomenological theory of superconductivity, we decided in 1946 to repeat Casimir's experiment, but before these plans had got very far Professor Casimir informed us that he had discovered a possible reason for the negative result of his original experiment. The mercury cylinder he had used was sealed off under vacuum in a quartz tube on which the coils were wound, and this tube would have slightly increased its diameter as the external pressure was reduced in order to lower the temperature of the helium bath. An estimate showed that this would have caused a slight increase of mutual inductance which roughly compensated the expected decrease due to the reduction of penetration depth. At about the same time further confirmation of the temperature variation

of penetration depth was obtained by Désirant & Shoenberg (1948), who measured the static magnetic properties of thin superconducting cylinders, and by Pippard (1947), who measured the difference in R.F. skin depth in the superconducting and normal states. In view of this new evidence, the emphasis of our plans changed somewhat; instead of investigating the cause of a negative result—as, for instance, by varying the frequency of measurement—we concentrated on using the method to obtain accurate measurements of the change of penetration depth with temperature in various conditions.

Positive results were indeed obtained with both tin and mercury, but it soon became apparent that the state of the surface of the specimen had to be carefully controlled before these results could be considered to have any intrinsic value. Early in the investigation it seemed very possible that the penetration depth was strongly dependent on the direction of current flow with respect to the crystal lattice (Laurmann & Shoenberg 1947), and most of the subsequent work was concerned with studying this possibility; it has turned out in fact that the anisotropy, if it exists at all, is much smaller than seemed at first indicated. Another question that has been investigated is the possible dependence of penetration depth on the magnitude of a steady magnetic field superimposed on the alternating measuring field; the indications of such an effect previously found (Désirant & Shoenberg 1948) seem to have been due to secondary causes, and the present experiments show that, up to 80 % of the critical field, the penetration depth does not increase by more than a few per cent.

DETAILS OF EXPERIMENTAL METHODS

The measuring apparatus

Since the apparatus was in most respects very similar to that already described by Casimir, only brief reference to quantitative details need be given, apart from a few points of difference in principle.

The detailed arrangement of the coil system is shown in figure 1; the purpose of the upper pair of coils, which are connected to give a mutual inductance opposite to the main pair, is to compensate most (usually more than 90 %) of the residual mutual inductance left over when a superconducting specimen is inserted as shown. This residual mutual inductance arises from the space occupied by the quartz tube and the gap between the specimen and the quartz, and is fairly large, about $1500\ \mu\text{H}$ for a 9 mm. diameter specimen. If the whole residual mutual inductance were compensated by a variable mutual inductance at room temperature, small changes of room temperature in the course of an experiment could (on account of thermal expansion) cause significant errors of order $0.01\ \mu\text{H}$, and special precautions would be necessary to keep the room temperature constant; by putting most of the compensation in liquid helium where thermal expansion effects are negligible this trouble is avoided. The quartz tube containing the coils was placed in the appendix of a Dewar vessel, which could be filled with liquid helium. It was found that slight changes of the inclination of the specimen to the quartz tube could cause appreciable changes of mutual inductance, and so to prevent rattling, the specimen was usually

lightly wedged with paper, the quartz tube was secured with cotton-wool plugs, and care was taken not to touch the apparatus during a series of measurements.

A Hartshorn bridge was used to measure the mutual inductance, as in Casimir's experiment. The bridge (figure 2) was fed by a phase-shift oscillator supplying alternating current of frequency 72.5 c./sec. and of good wave form. The variable mutual inductance was a Tinsley standard, which could be varied in steps of $10\ \mu\text{H}$, in series with a continuously variable mutual inductance, designed by Dr Ashmead, which could be set and read reliably to $0.01\ \mu\text{H}$. The output of the bridge was taken through a three-stage amplifier to a Campbell vibration galvanometer.

Measurements of the in-phase component M' of the mutual inductance proved to be very important, and could be deduced from the resistance R of figure 2 by the relation

$$M' = rr' / \omega(R + r + r'), \quad (1)$$

where ω is the angular frequency. For small values of M' , the balance was not appreciably disturbed until R differed considerably from the true setting, and so a 'bracketing' method was used to find M' more accurately. A value of R was found for which the galvanometer deflexion was equal to that for the switch S open (R infinite); the true value of R was then just twice as great (since for large R , r and r' may be neglected in the denominator).

A good deal of attention had to be paid in the lay-out of the apparatus to screening connexions from the low-temperature part of the apparatus to the bridge and to earthing the right points of the apparatus, to avoid

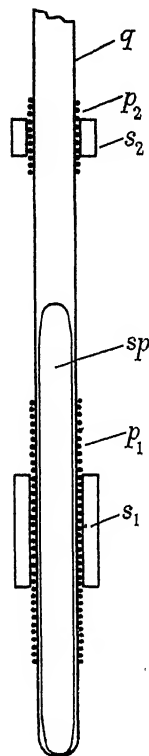


FIGURE 1. The coil system (half actual size). p_1, p_2 , main and compensating primaries 62.8 turns per cm.; s_1 , main secondary 8000 turns; s_2 , compensating secondary 4000 turns; q , quartz tube; sp , specimen.

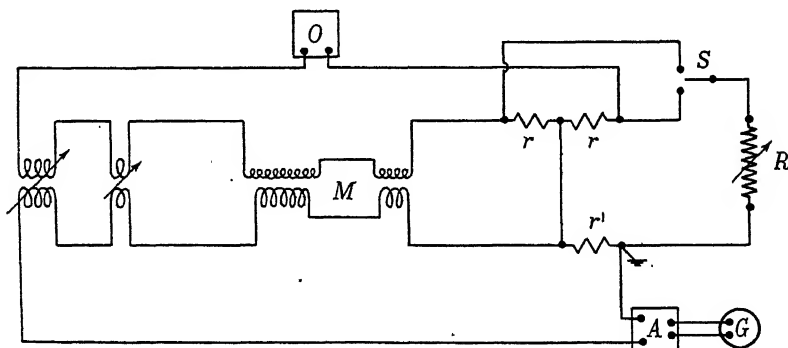


FIGURE 2. The measuring circuit. O , oscillator; A , amplifier; G , vibration galvanometer; S , two-way switch to change from positive to negative in-phase adjustment; R , six-dial resistance box; rr , fixed low resistances, each $1\ \Omega$; r' , fixed low resistance, $1.04\ \Omega$; M , the coil system shown in figure 1.

various stray capacity effects. The apparatus was very sensitive to local disturbances; some were caused by magnetic pick-up from neighbouring power cables, the band width of the detecting arrangements not being small enough to cut out 50 c./sec. effects completely, and some (of an irregular kind) due to some complicated effect of high-frequency radiation, as, for instance, from the neighbouring cyclotron. Most of these troubles were avoided by working at night when the disturbances were usually found to be very much smaller.

The overall sensitivity was such that for 20 mA in the primary circuit (which gave a peak field of about 2 gauss at the surface of the superconductor) a change of $1 \mu\text{H}$ mutual inductance gave a deflexion of 8 cm. of the vibration galvanometer; since in good conditions the 'zero' deflexion (i.e. for no current in the bridge) was less than 1 mm., it was possible to detect changes of mutual inductance of less than $0.01 \mu\text{H}$, i.e. somewhat beyond the accuracy with which the continuously variable mutual inductance could be read. In practice usually the mean of five readings was taken for each measurement and, as will be seen later (p. 575), the consistency of the experimental results indicated that such means had standard deviations of order $0.004 \mu\text{H}$.

Method of reduction of observations

The mutual inductance of the coil system when a superconducting rod of cross-section A and perimeter p is inserted can be written as

$$M = M_0 - k(A - \lambda p) - m_0, \quad (2)$$

where M_0 is the mutual inductance of the empty bottom pair of coils, m_0 that of the top compensating pair (which, as was verified experimentally, was sufficiently remote to be unaffected by the presence of the superconductor), k is a constant and λ is the penetration depth of a magnetic field into the superconductor. Since the radius of the superconductor is much larger than the penetration depth, the latter can be defined, independently of any assumptions about the law of penetration, as

$$\lambda = \frac{1}{pH_0} \int_A H dS, \quad (3)$$

where H_0 is the applied field, and H is the field inside the superconductor. If now the temperature is varied, the only changes in M , apart from possible small corrections which will be considered later (see p. 564), arise from changes in λ , and we have

$$\Delta M = M(T) - M(T_0) = kp[\lambda(T) - \lambda(T_0)] = kp\Delta\lambda, \quad (4)$$

where T_0 is some standard temperature, usually 2.17°K in our experiments.* Thus if k and p are known, $\Delta\lambda$ can be found from experimentally observed values of ΔM . If the primary coil and the superconducting rod can be treated as infinitely long,

$$M_0 = k_0 A_0, \quad (5)$$

$$k = k_0 = 4\pi nN, \quad (6)$$

where n is the number of turns per unit length, and A_0 the cross-section of the primary coil and N the total number of turns of the secondary.

* Since λ varies very little with temperature far below the transition temperature, the precise value of T_0 is not important.

In fact, since the primary coil is of finite length, k may be appreciably less than k_0 , and its value was found experimentally both by measuring the decrease of M when a superconductor of known cross-section area A was inserted, and from the slope of the straight line obtained when M is plotted against A for the 20 % range of A covered by the different specimens used. These values of k agreed to better than 1 %, and were about 8 % lower than k_0 . Another estimate could be made from the measured absolute value of M_0 , using (5), and although it is not obvious that this value of k should be quite identical with that used in (2) it did in fact agree to 1 % with the previous estimates, and also agreed reasonably with a calculation of M_0 based on the detailed dimensions of the coils.

In applying (4) to calculate $\Delta\lambda$, we shall assume that $p = 2\pi r$, where r is the radius of the specimen. It will be convenient to postpone a discussion of the validity of this assumption, equivalent to assuming a perfectly smooth surface, until the experimental results are considered. The value of r for the tin specimens was measured directly by a micrometer, but for the mercury specimens this could not be done so easily, and A (and hence r) was deduced from the value of M at T_0 , using equation (2), and the experimentally determined value of k .

Discussion of possible systematic errors

Since the measured effect is so small it is necessary to consider various possible causes of systematic error before we can have confidence that observed changes of M are not due to causes other than the changes of penetration depth we wish to measure.

One such cause might be a change of specimen radius due to thermal contraction. The coefficients of thermal expansion of tin and mercury at liquid helium temperatures have not been measured, but, on the basis of Grüneisen's law, can be estimated. It turns out that the lowering of temperature from 4 to 2° K could produce a change of radius of at most 10^{-7} cm. (in mercury, for which the coefficient of thermal expansion is greatest). Since this is only just comparable with the accuracy to which $\Delta\lambda$ can be measured, no correction has been made for this effect.

A slightly larger effect comes from change of radius caused by the change in hydrostatic pressure of the helium bath as the temperature is lowered; this can be calculated fairly accurately (using data of Grüneisen & Skell (1934) for solid mercury) and the appropriate small corrections (of order 2×10^{-7} cm.) have been applied in the results.

A number of other possible sources of error were eliminated by a blank experiment in which the specimen was removed and it was established that M changed by less than $0.01 \mu\text{H}$ when T was reduced from 4.2 to 2.1° K. This showed that no systematic error greater than 0.6×10^{-8} cm. could arise from any changes in the coil system, such as thermal contraction, mechanical changes, or changes of magnetic properties, associated with the reduction of temperature. One cause of systematic error, due to the spreading of the superconducting transition somewhat below the transition temperature, will become apparent when the actual experiments are described, and is discussed later (p. 571).

Temperature measurement

The temperatures were deduced from the vapour pressure over the helium bath using the 1932 Leiden scale (Keesom 1932); although this scale may be in error by as much as 0.01°K in some regions, this is unimportant, since for our purposes it is only a high relative accuracy that is required. No stirrer was used in the bath, but it was found that temperature equilibrium was quickly achieved during lowering of temperature by reducing the pressure. Rapid warming was produced by using the main secondary coil as an electric heater, but thermal equilibrium was only slowly achieved, so all our readings were taken during cooling.

Preparation of specimens

The specimens used at liquid helium temperatures are listed in table 1* together with such relevant features as will be needed later.

TABLE 1

specimen	material	diameter (mm.)	prepara- tion method	crystal orientation (ψ)	ρ at 18°C for tin; at 80°K for mercury ($\Omega\text{cm.} \times 10^6$)	ρ at 4.2°K ($\Omega\text{cm.} \times 10^9$)
T 0	12966	8.67	<i>a</i>	—	—	2.5
T 1	2135	9.16	<i>b</i>	15°	10.3	3.9
T 3	2135	9.16	<i>b</i>	79°	12.7	4.1
T 5	2135	9.26	<i>b</i>	90°	12.7	4.4
T 6	2356	8.60	<i>b</i>	1°	10.4	1.9
T 9	2356	9.45	<i>b</i>	8°	10.4	2.1
T 10	2135	8.98	<i>f</i>	{ several crystals all $\sim 90^\circ$	13.8	4.0
M 2	H 11023	9.86	<i>c</i>	—	—	34.3
M 3	H 11023	9.86	<i>c</i>	—	—	31.9
M 4	H 11023	9.86	<i>c</i>	—	6.57	42.5
M 7	2505	7.76	<i>d</i>	—	6.63	38.5
M 8	2505	7.95	<i>d</i>	—	6.62	44.0
M 9	2505	7.58	<i>d</i>	—	6.33	40.4
M 13	2505	9.00	<i>e</i>	33° (<i>e</i>)	6.79	40.2
M 14	2505	9.00	<i>e</i>	62° (<i>e</i>)	6.38	37.0
M 15	2505	9.00	<i>e</i>	{ at least 3 crystals $89^\circ, 117^\circ, 65^\circ$	6.03	35.2
M 16	2505	9.00	<i>e</i>	$27^\circ, 31^\circ$ (<i>e</i>)	6.82	40.0
M 17	2505	9.00	<i>e</i>	$56^\circ, 65^\circ$ (<i>e</i>)	6.34	37.0

Notes. T = tin, M = mercury. Material was all spectroscopically pure, Johnson Matthey except where marked H (Hilger H.S. quality). Where no crystal orientation is given it is probable that specimen was polycrystalline. Orientations deduced electrically are marked (*e*).

The various methods of preparation were as follows:

(a) Cast in a glass tube under vacuum; cracked away from the glass and finally carefully machined.

* Only those specimens to which reference is made in the discussion were tested. Actually twenty-five specimens in all were used in forty liquid helium experiments.

(b) Deliberately grown as a single crystal in a glass tube under vacuum by slowly lowering a surrounding furnace; a seed crystal of approximately the desired orientation but much smaller diameter was used in each case. When cold, the glass wall was broken away as carefully as possible and the seed and tailpiece cut off with cutting pliers; it was noticed that the glass often cracked during the slow-cooling process. This was probably due to the sticking of the metal to the walls, causing implosion as the tin contracted on cooling.

(c) Mercury slowly cooled from the bottom in the quartz tube on which the primary coils were wound. Except on one occasion, the quartz tube was found broken at the end of the experiment each time this method was used, probably for the same reason as mentioned above for tin. It is possible that the breakage may have been due to expansion of the mercury on re-melting, but, since care was taken to avoid this by warming from the open end, the 'implosion' explanation is more probable. These breakages were of course very awkward, since they necessitated making a new mutual inductance on each occasion, and eventually the following two methods were tried.

(d) Mercury slowly cooled from the bottom in a separate quartz tube of smaller diameter which could be lowered into the quartz tube on which the primary coils were wound. As will be seen later, the results obtained by this method, even though considerable variations of the cooling schedule were tried, were mostly unsatisfactory. In nearly every case the quartz tube was found broken at the end of the experiment; in one or two cases when it was possible to examine the tube immediately after removal from the helium bath, but before the mercury had warmed up to liquid nitrogen temperatures, the tube was found already cracked, showing clearly that the breakage occurred during cooling rather than warming. It was this evidence that eventually led to the 'implosion' explanation, and, as will be seen later (p. 576), the same phenomenon of the mercury sticking to the quartz is probably responsible for the unsatisfactory results obtained with these specimens.

(e) Free-mercury single-crystal rods were prepared by the method of Andrade & Hutchings (1935) in which liquid mercury in a precision-bore glass tube whose internal wall has been wetted with alcohol, and closed by a steel plug at the bottom, is slowly lowered into alcohol cooled by solid CO_2 .^{*} After solidification, the steel plug (coned to a point internally to promote growth of a single crystal) was withdrawn and usually the mercury rod was found to slide out of the tube quite easily. Before inserting the rod into the experimental quartz tube (which was pre-cooled to 80° K), excess alcohol was wiped off with a rag to prevent sticking, and the rod cooled in liquid air (to avoid the danger of melting during the transfer process); the rod was handled by a long quartz rod enlarged at its bottom end which was embedded in the mercury during solidification. It is of interest to note that three of these rods

^{*} Owing to the larger length and diameter of our rods, some slight modifications were found necessary. The lubricating alcohol had to be made rather more viscous by the addition of glycerine and the cooling had to be completed rather rapidly; it seemed that if the alcohol was too 'thin' or the cooling too slow, the lubricating alcohol was able to drain away before the mercury solidified, and thus did not prevent sticking to the glass. The rapid cooling may have been responsible for the fact that one of our rods proved to consist of several rather than a single crystal.

(M 15, 16 and 17) were preserved for 3 months in liquid air while arrangements for X-ray examination were being prepared.

(f) Attempts were made to grow tin crystals using a technique similar to that outlined in (e). The steel plug was replaced by a similar carbon plug, and the lubricant used was a silicone oil instead of alcohol. It was hoped that by avoiding sticking to the walls a slightly better quality specimen might be obtained. In fact the results were no better than those obtained by (b), and, moreover, the specimen actually used proved later to be not a single crystal.

Examination of the specimens

(a) Visual

Parts of the surface of the tin crystals were covered with small hemispherical pits caused either by bubbles or by shrinkage of the tin during solidification. A rough estimate showed that although these pits were numerous, they covered only a very small fraction of the surface area, and so their effect in increasing the perimeter was probably negligible. The fact that both the tin and mercury crystals appeared highly polished suggested that there were no irregularities on a microscopic scale (i.e. comparable with a wave-length of light) which might increase the perimeter. The etched specimens could be seen to be very rough, and it was not surprising that the effective perimeter of these proved to be much greater than $2\pi r$.

(b) Electrical

Measurements of specific resistance at low temperatures were useful as a clue to the quality of the specimens, and at high temperatures as indications of crystal orientation (see below). The values of ρ given in table 1 were deduced from the changes of mutual inductance caused by eddy currents in the specimen, using standard formulae for the complex permeability of a cylinder in an alternating magnetic field. For tin ρ at 4.2°K is mostly 'residual' resistance, and the low values found confirm the high purity of the material used. It is interesting that the values of ρ at 4.2°K seem to be correlated strongly with the batch of tin used, but hardly at all with crystal orientation, suggesting that perhaps residual resistance is not as strongly anisotropic as ideal resistance.

For mercury, ρ at 4.2°K is almost entirely 'ideal' resistance, and there was an interesting contrast in the behaviour of the 'good' specimens (prepared by (e) above) and the others. For 'good' specimens the values of ρ were in the same ratios at 4.2°K as at 80°K , showing that the same arrangement of crystals persisted near the surface (within the skin depth of order 0.6 mm.) as in the body of the specimens. For most of the 'poor' specimens, however, there seemed to be no correlation between the two sets of resistance values, suggesting either that strains at the surface have upset the crystal orientations there, or more probably that these strains have caused the 'residual' resistance to become so large as to be comparable with the 'ideal' resistance.

(c) Crystal orientation

For the tin crystals, the angle ψ between the tetragonal axis and the rod axis was found to within a few degrees by the etch-pit reflexion method, using either the etched specimen itself or the etched tailpiece. The specific electrical resistance ρ at room

temperature provided useful confirmation of these determinations, the ratio of the values for the 'parallel' and 'perpendicular' crystals being in good agreement with that found by Bridgeman (1925).^{*} Some electron diffraction studies by Mr J. W. Menter on tin specimens similar to those used in the low-temperature experiments confirmed that they were single crystals within the small depth of penetration of the electrons used ($\sim 5 \times 10^{-7}$ cm.). This eliminates the possibility that the surface layer in which the superconducting current flows has a structure differing from the bulk of the specimen.

For mercury, the orientations were deduced from the value of ρ at 80° K. It can be shown (Fraser & Shoenberg 1949) that if the principal axis of a single-crystal cylinder is inclined at angle ψ to the cylinder axis, the cylinder behaves in an alternating field as if it had an isotropic resistivity ρ given by

$$\rho = \frac{1}{2}[(\rho_1 + \rho_3) + (\rho_1 - \rho_3) \cos^2 \psi], \quad (7)$$

where ρ_1 and ρ_3 are the resistivities perpendicular and parallel to the principal axis. Thus, using Sckell's (1930) values, $\rho_1 = 7.04 \times 10^{-6} \Omega \text{ cm.}$, $\rho_3 = 5.35 \times 10^{-6} \Omega \text{ cm.}$ at 80° K, the orientation ψ of a cylinder can be estimated from its value of ρ . For M 15, 16 and 17 back-reflexion Laue X-ray studies at liquid air temperature were made by Mr J. V. Smith, and the orientations found for M 16 and 17 agreed within experimental accuracy with those deduced electrically; M 15 (in agreement with visual indications of grain boundaries) proved to consist of at least three crystals with their principal axes roughly perpendicular to the rod. It should be noticed that (7) applies only to a single crystal; if the specimen contains more than one crystal, ρ may have any value between ρ_1 and ρ_3 , and in such cases ρ gives a measure of the degree of preferred orientation. For instance, $\rho = \rho_3$ would indicate that all the crystals had their trigonal axes perpendicular to the cylinder axis, and also perpendicular to a radius normal to the cylinder axis.

EXPERIMENTAL RESULTS

General features

Measurements of M and M' (the real and imaginary parts of the mutual inductance) were made as the temperature was lowered, and typical curves are shown in figures 3, 4 and 5. In the normal state M and M' are nearly constant, and from their values the specific resistance at 4.2° K was deduced. During the transition to superconductivity, both M and M' fall rapidly over a narrow temperature interval—of order 0.01° K for a good tin specimen, 0.003° K for a 'good' mercury specimen, and 0.02° K for a 'poor' mercury specimen. The transition temperature T_c may be taken with an uncertainty usually less than 0.001° K as the temperature at which M has completed half its descent.

For further lowering of temperature, the variation of M and M' is very much less rapid and is shown in figures 6, 7 and 8, where the scales of ordinates are greatly

^{*} Our absolute values were, however, about 5 % higher than those of Bridgeman; part of this discrepancy may be due to slight systematic errors in our method of finding ρ .

expanded (400 times in figure 6 and 1000 times in figures 7 and 8). In these diagrams it is convenient to plot not M and M' , but their differences ΔM and $\Delta M'$ from the values they assume at a low temperature (2.17°K). In fact M' should become zero

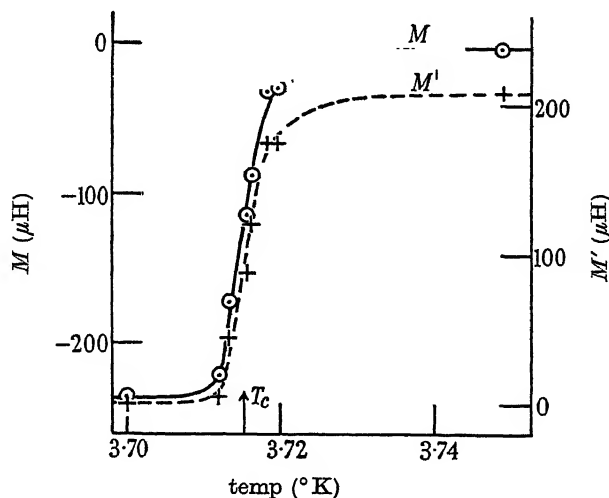


FIGURE 3. Transition to superconductivity of T 9.

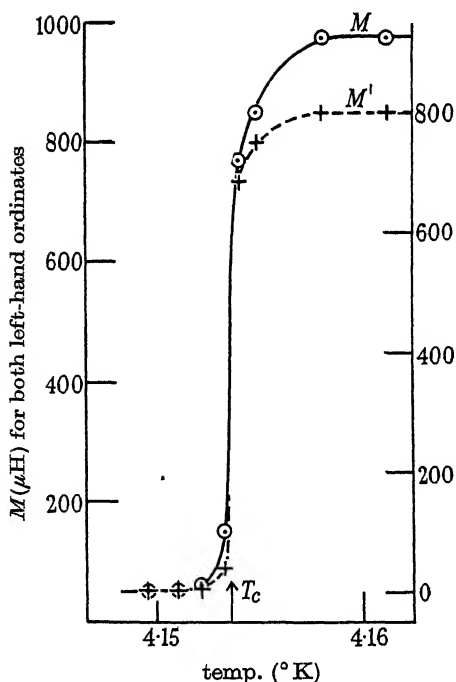


FIGURE 4. Transition to superconductivity of a 'good' mercury specimen, M 16.

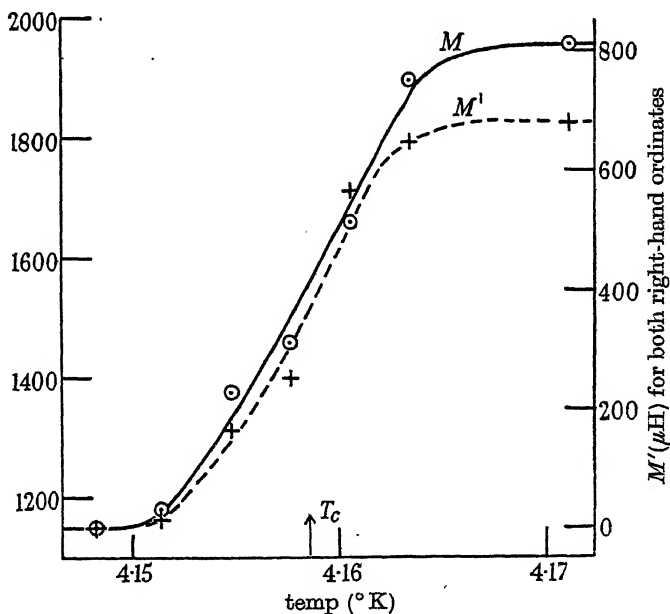


FIGURE 5. Transition to superconductivity of a 'poor' mercury specimen, M 9.

if no energy is dissipated in the specimen, but owing to small dissipation elsewhere in the circuit, M' approaches a constant value (usually of order $0.2\mu\text{H}$) at low temperatures, and it is reasonable to assume that only $\Delta M'$ represents the dissipation

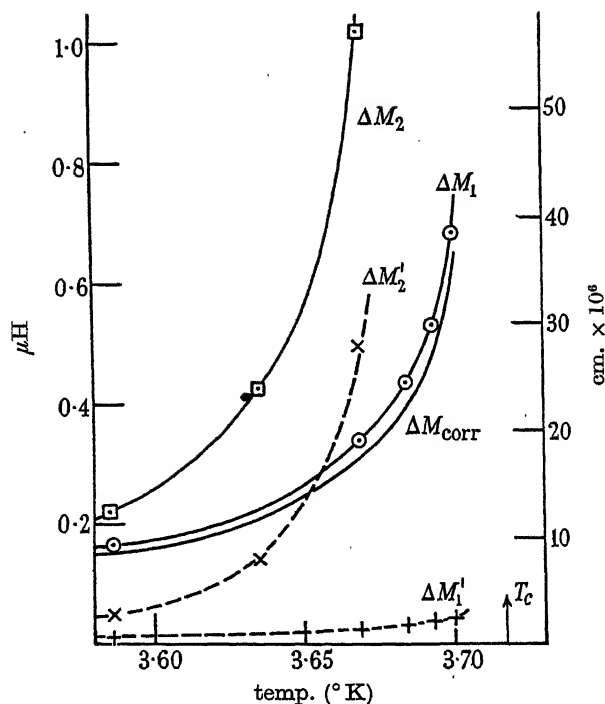


FIGURE 6. Variation of ΔM and $\Delta M'$ for T9 below T_c , illustrating method of correction for in-phase effect. ΔM_1 , $\Delta M'_1$, with earth's field compensated; ΔM_2 , $\Delta M'_2$, with earth's horizontal field not compensated; ΔM_{corr} , corrected for in-phase effect.

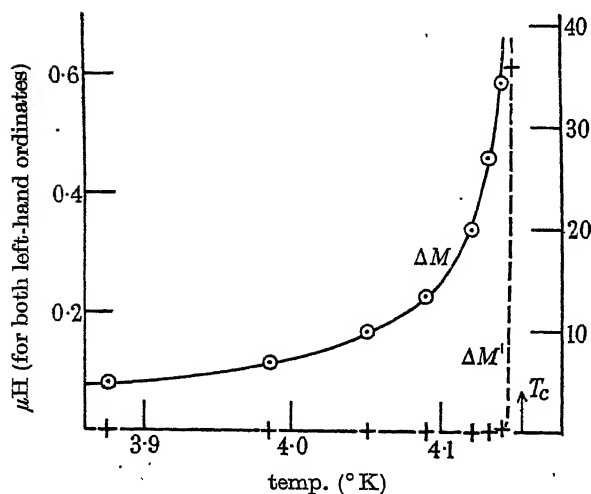


FIGURE 7. Variation of ΔM and $\Delta M'$ for a 'good' mercury specimen, M 15, below T_c (earth's field compensated). No in-phase correction is necessary.

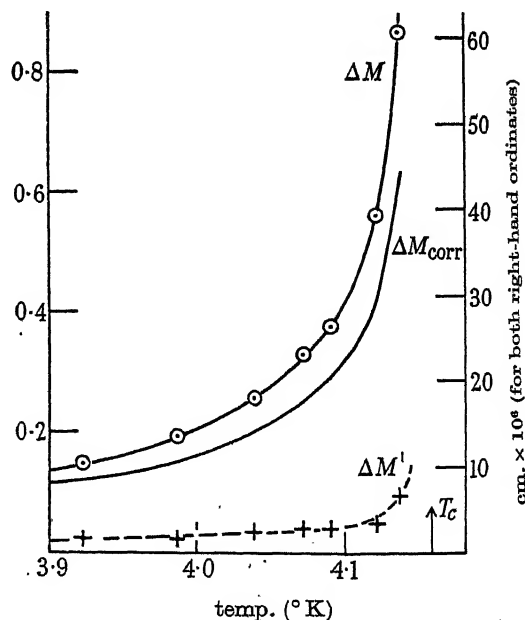


FIGURE 8. Variation of ΔM and $\Delta M'$ for a 'poor' mercury specimen, M 9, below T_c (earth's field compensated). ΔM_{corr} , corrected for the in phase effect.*

* The curves for the earth's field uncompensated cannot be shown conveniently on the same scale: thus at the highest temperature, 4.134°K , $\Delta M_2 = 5.86 \mu\text{H}$ and $\Delta M'_2 = 1.92 \mu\text{H}$, while at 4.071°K , $\Delta M_2 = 1.53 \mu\text{H}$ and $\Delta M'_2 = 0.67 \mu\text{H}$.

effects due to the specimen. As for ΔM , this should be proportional to $\Delta\lambda$, as already explained, and, as can be seen from figures 6, 7 and 8 the changes expressed in cm. by means of (4) are of the expected order of magnitude.

It will be seen that for M15, $\Delta M'$ has fallen to practically zero within about 0.01°K of T_c , but for T9 and more so for M9, $\Delta M'$ is still appreciable at much lower temperatures. The probable interpretation of a non-zero $\Delta M'$ is that a very small and diminishing fraction (of order 1 in 5000 for T9) of the metal surface remains in the normal state below T_c , and that $\Delta M'$ arises from eddy-current losses in these normal regions. In other words, although most of the transition to superconductivity takes place in a narrow range of temperature, the transition actually extends a long way below this range. Now in a normal region the alternating magnetic field will penetrate very much deeper than in a superconducting region, in fact, to the order of the skin depth, which for the frequency used is about 0.2 mm. in tin and 0.6 mm. in mercury. Thus the presence of a non-zero $\Delta M'$ is an indication that the ΔM may be due not only to change of genuine superconducting penetration depth but partly also to ordinary skin-effect penetration in small normal regions. That this is really so is shown by the fact that ΔM at a given temperature was larger if $\Delta M'$ was larger.

Particularly large $\Delta M'$ values can be obtained, as already noted by Casimir, if a specimen is cooled in the earth's magnetic field. The curves marked ΔM_2 and $\Delta M'_2$ in figure 6 refer to such a case, and it is evident that the part of ΔM_2 associated with the in-phase component $\Delta M'_2$ is quite comparable to the part with which we are concerned, arising from change of penetration depth. The large values of ΔM_2 and $\Delta M'_2$ are presumably due to 'freezing in' of the earth's horizontal field in a few patches where there are holes or inhomogeneities; as the metal cools, the critical field rises and the normal patches where the trapped flux issues shrink to keep this flux constant.

To avoid this effect, the earth's field was carefully compensated (to within about 1 %) in all the basic experiments (as, for instance, in the ΔM_1 and $\Delta M'_1$ curves of figure 6), but a subsidiary set of readings was also taken during a second cooling in which the earth's horizontal field was uncompensated (e.g. the ΔM_2 and $\Delta M'_2$ curves of figure 6). On the basis of these readings it was possible to estimate a correction to allow for that part of ΔM_1 associated with the in-phase component $\Delta M'_1$ and due to penetration into normal regions. Thus if it is assumed that this unwanted extra part is proportional to $\Delta M'_1$ (and it was found that this was roughly true), the extra part is given by

$$\Delta M'_1(\Delta M_2 - \Delta M_1)/(\Delta M'_2 - \Delta M'_1).$$

For T9, as shown in figure 6, this produces a correction of order 10 %, and the curve $\Delta M_{\text{corr.}}$ is obtained when this unwanted part is subtracted from ΔM_1 . This procedure was applied in all the results quoted below, but, since the accuracy of the correction is somewhat uncertain, it is not impossible that systematic errors of as much as half the correction may still be left, and where (as for M9) a correction of much more than 20 % is required, the results must be considered unreliable.

The 'pathology' of this 'in-phase' effect proved to be complicated, but it seems that the residual $\Delta M'$ effect when the earth's field has been compensated is least in

specimens with the least internal stresses. Thus there was a great difference between the mercury specimens prepared by the methods (d) (M 5 to 12) and (e) (M 13 to 17) above; the cracking of the quartz containers suggested that the former were under considerable stress and the $\Delta M'$ values for most of these were high, quite comparable occasionally with the high values obtained when the earth's field was uncompensated. The free rods, however, of which figure 7 represents a typical example, had almost negligible values of $\Delta M'$ except very close to T_c . As already mentioned the transition curves for the bad specimens were much broader than those for good specimens, again suggesting a worse state of internal stress. A further indication that $\Delta M'$ is associated with some sort of defect is that generally (though not always) the $\Delta M'$ values with the earth's field compensated are highest for specimens where lack of compensation during cooling has the biggest effect in increasing $\Delta M'$; thus for some of the free mercury rods where $\Delta M'$ was almost negligible the earth's field had hardly any effect either, except very close to T_c .

For any one specimen, the $\Delta M'$ values were not always reproducible from one experiment to another, and sometimes even within the same liquid helium run the $\Delta M'$ values became twice as large (and the ΔM values increased correspondingly) in a second cooling, although the earth's field was compensated throughout.* Occasionally it was found possible to reduce $\Delta M'$ considerably by temporarily increasing the measuring current enough to start destruction of superconductivity. These various effects suggest that the normal regions responsible for $\Delta M'$ are not very stable, and that their formation and arrangement may depend on rather trivial factors, but no detailed explanation of all the observed peculiarities has been found.

In the above description no mention has yet been made of the importance of the measuring current i . Where $\Delta M'$ was negligible, ΔM was accurately independent of i , but where $\Delta M'$ was appreciable, both ΔM and $\Delta M'$ usually increased with i , presumably due to the magnetic field of the current increasing the fraction of normal metal present. When the correction for $\Delta M'$ was not too big, it was found that the corrected ΔM was usually practically independent of i , which gave confidence in the reliability of the correction, and our procedure was to use the highest value of i which did not cause too big a value of $\Delta M'$. For about 0.01°K below T_c , i had to be less than 5 mA, and it was usually impossible to get reliable readings of M and M' closer than this to T_c , both because only poor accuracy could be obtained with values of i below 5 mA, and because, owing to spread of the transition, $\Delta M'$ was nearly always large in this region. For lower temperatures the value of i could be increased, and usually it was possible to use 20 mA for temperatures more than 0.02°K below T_c ; there was usually no gain in accuracy in going beyond 20 mA.

Quantitative data

A direct plot of $\Delta\lambda$ against T as in figures 6 and 7 does not lend itself conveniently to quantitative discussion because of the very steep rise of $\Delta\lambda$ close to T_c . The nature of the results is better displayed if use is made of the relation

$$(\lambda/\lambda_0)^2 = 1/(1 - (T/T_c)^4), \quad (8)$$

* Our preliminary observation that long exposure to air systematically increases $\Delta M'$ for tin, has not been confirmed.

which fits the mercury colloid results well (Shoenberg 1940; Daunt, Miller, Pippard & Shoenberg 1948). From this relation it follows that

$$\Delta\lambda = \lambda_0(1 - (T/T_c)^4)^{-\frac{1}{2}} - \lambda(2.17^\circ \text{K}), \quad (9)$$

so it is appropriate to plot experimental values of $\Delta\lambda$ against z , where

$$z = (1 - (T/T_c)^4)^{-\frac{1}{2}}, \quad (10)$$

rather than against T itself.

Good linear plots are in fact obtained, as shown in figures 9 and 10, for some examples which are discussed below; the intercept on the $\Delta\lambda$ axis gives λ (2.17°K) and the slope gives λ_0 , the penetration depth at 0°K , if the relation (8) is true. This proviso is necessary, since in fact most of the experimental points are for T close to T_c , and an almost equally good straight line (but of different slope) would be obtained if the 4 in (8) were replaced by 3 or 5, so that (8) cannot be regarded as more than roughly confirmed by our experiments.

All the experiments have been analyzed by this method and the results are summarized in table 2. Usually a least-squares analysis has been made and the slope λ_0 and intercept, λ (2.17°K), are given together with their standard deviations.

TABLE 2

specimen	date of experiment	T_c ($^\circ \text{K}$)	λ_0 (cm. $\times 10^6$)	λ (2.1°K) (cm. $\times 10^6$)	μ (cm. $\times 10^6$)	highest value of z used	in-phase correction at highest z (%)
T 0	18. vii. 47	3.712	5.60 ± 0.07	5.4 ± 0.2	0.34	6.24	2
T 1	28. i. 48	3.714	5.11 ± 0.06	5.8 ± 0.2	0.23	5.26	2
T 1 etched	11. ii. 48	3.715	17.2	—	—	6.67	0
T 1 polished	26. xi. 48	3.715	5.19 ± 0.05	5.1 ± 0.2	0.32	7.65	1
T 6	27. ii. 48	3.715	5.32 ± 0.09	5.8 ± 0.4	0.50	6.39	10
T 9	5. iii. 48	3.716	5.34 ± 0.07	5.5 ± 0.2	0.29	5.50	9
T 9	10. iii. 48	3.717	5.13 ± 0.05	5.4 ± 0.2	0.20	5.34	9
T 9	12. iii. 48	3.717	5.03 ± 0.10	5.4 ± 0.3	0.42	5.25	7
T 3	4. ii. 48	3.715	5.63 ± 0.06	6.0 ± 0.2	0.23	5.13	10
T 3 polished	3. xii. 48	3.716	5.28 ± 0.08	5.3 ± 0.2	0.24	4.53	11
T 5	18. ii. 48	3.715	5.67 ± 0.13	5.8 ± 0.4	0.41	4.53	2
T 5	17. iii. 48	3.715	5.77 ± 0.05	6.0 ± 0.2	0.17	5.50	8
T 5	19. xi. 48	3.715	5.78 ± 0.05	5.8 ± 0.2	0.33	7.65	7
T 5 polished	10. xii. 48	3.716	4.91 ± 0.08	4.4 ± 0.3	0.49	7.42	13
T 10	5. xi. 48	3.717	5.5	—	—	4.53	11
M 2	25. vii. 47	4.156	10.0	—	—	6.75	33
M 3	8. viii. 47	4.158	4.10 ± 0.05	4.5 ± 0.2	0.28	6.50	9
M 4	14. v. 48	4.160	5.5	—	—	4.92	12
M 7	2. vi. 48	4.154	4.06 ± 0.11	4.6 ± 0.5	0.41	4.85	13
M 8	4. vi. 48	4.164	17.0	—	—	3.80	11
M 9	9. vi. 48	4.158	7.2	—	—	6.90	31
M 13	18. viii. 48	4.154	4.48 ± 0.09	5.5 ± 0.3	0.51	7.42	0
M 14	20. viii. 48	4.156	4.33 ± 0.07	5.5 ± 0.3	0.47	8.36	3
M 15	1. x. 48	4.151	4.08 ± 0.02	4.2 ± 0.1	0.14	7.83	2
M 16	15. x. 48	4.153	4.34 ± 0.04	4.5 ± 0.3	0.22	8.64	0
M 17	10. xi. 48	4.153	4.22 ± 0.03	4.2 ± 0.1	0.18	8.98	1

Notes. No values of the intercept λ (2.1°K) are given where only graphical analysis has been made, since the accuracy would be poor. The percentage corrections quoted are the percentages of the original ΔM values, which have been subtracted.

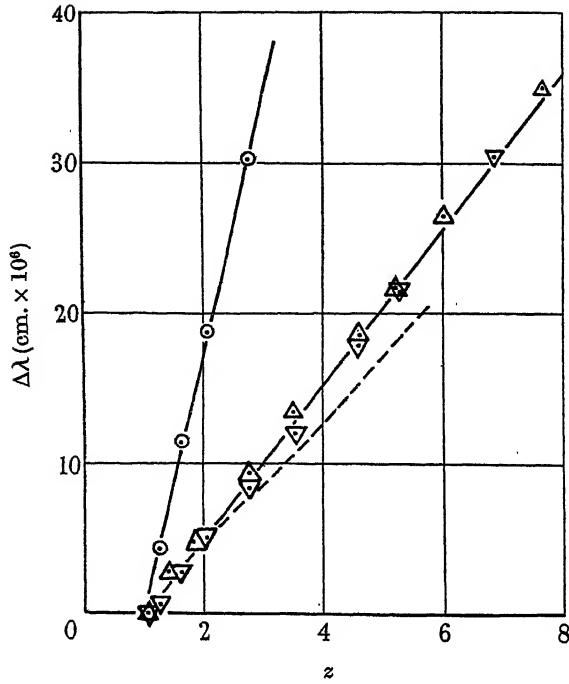


FIGURE 9. Variation of $\Delta\lambda$ with z for tin. ∇ Tl as cast; \odot Tl after acid etching; \triangle Tl after electrolytic polishing. The broken curve is based on the results of Pippard (1947).

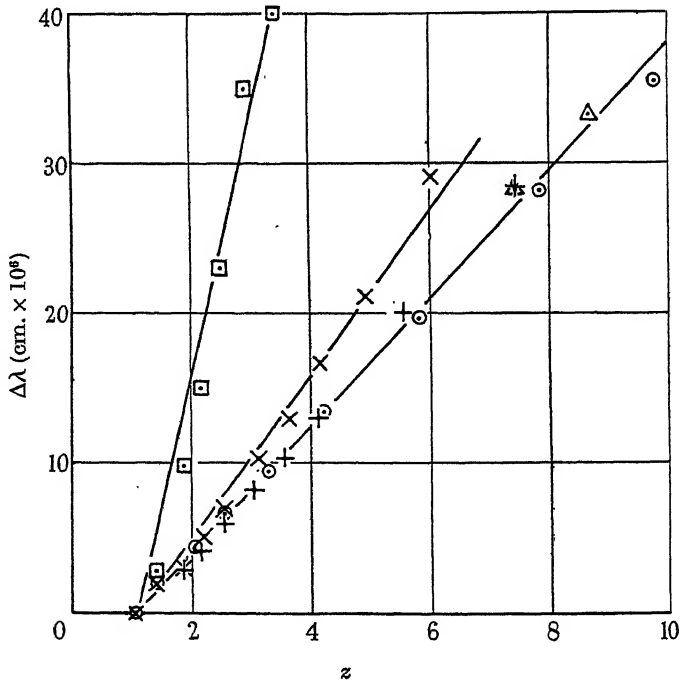


FIGURE 10. Variation of $\Delta\lambda$ with z for some mercury specimens. \square M8; \times M4; \odot M15; $+$ M13; \triangle M16. To avoid confusion, some of the experimental points have been omitted.

Where the results are required only for qualitative discussion, only the slope obtained from a graphical plot is quoted and no standard deviation is given. Before making a least-squares analysis each set of data was plotted graphically, and in some cases a few points for very high z , which lay obviously off the straight line plot, were rejected. Such departures could be due either to small errors in T or T_c (for instance, 3 % error in z at $z = 7$ is caused by 0.001°K error in T or T_c) or to inaccurate estimation of the in-phase correction, which usually increases with z (the obvious departures from linearity were usually most marked in experiments where this correction was large). For each series we quote the highest value of z used in the analysis, and the size of the in-phase correction at that point, which as already mentioned gives an idea of the possible systematic error.

DISCUSSION OF RESULTS

As can be seen from table 2, the slope estimates of λ_0 are much more reliable than the intercept estimates of λ (2.17°K), and so, apart from noting that the estimates of λ (2.17°K) are of the right order of magnitude and usually reasonably close to $\lambda_0(1 - (2.17/T_c)^4)^{-\frac{1}{2}}$ (this is $1.04\lambda_0$ for mercury and $1.06\lambda_0$ for tin), we shall make no further use of the intercepts.* The values of μ , the standard deviation of any single point from the linear plot, are surprisingly low, ranging from 15 to 50 Å; these correspond to from 0.0025 to 0.0085 μH expressed in mutual inductance, and as already mentioned (p. 563) are slightly lower than the precision with which any individual setting can be read. Where any specimen has been measured on more than one occasion even after a long interval the consistency is reasonably satisfactory (e.g. see results for T 5 and T 9), which gives added confidence in the significance of the results.

In our early experiments the large difference in λ_0 between M 2 and M 3 suggested that perhaps these two specimens had different crystal orientations and that λ_0 depended strongly on the direction of the superconducting current relative to the crystal axes (Laurmann & Shoenberg 1947).† Most of the subsequent experiments were directed to investigating this suggestion, and it soon became very unlikely that all the difference between M 2 and M 3 could be due to such an anisotropic effect. For instance, M 4, 7 and 8 all had approximately the same specific resistance at 80°K and so presumably the same kind of arrangement of crystals, and yet the values of λ_0 are very different. It became in fact much more likely that these differences and also the originally noted difference between M 2 and M 3 were due to surface conditions. More direct evidence of the lack of any marked anisotropy was eventually obtained from the 'good' specimens M 13 to 17, which although clearly varying in crystal orientation show nearly equal values of λ_0 .

The importance of surface conditions was strikingly brought out by the behaviour of T 1 after acid etching. As can be seen from figure 9 and table 2, the apparent penetration depth goes up by a factor of 3.4, presumably due to roughening of the surface, which effectively increases the perimeter round which the superconducting current

* A bad reading at 2.17°K can cause a large error in the intercept without much affecting the slope; this is probably the cause of the high intercepts for M 13 and 14.

† In the preliminary note T_c for M 2 was wrongly given as 4.165°K .

flows. That the magnification is due entirely to surface conditions was further demonstrated by the fact that the original results were reproduced when the surface was made smooth again by electrolytic polishing. It thus seems probable that the anomalously large values of λ_0 found with some of the mercury specimens are due to cracking of the surface associated with sticking of the mercury to the quartz container as it cools (as already mentioned on p. 566).

It is possible that surface conditions are responsible for a misleading appearance of anisotropy for the tin crystals too. Thus our first experiments indicated a difference between the results for the 'perpendicular' crystals T3 and T5 (average $\lambda_0 = 5.71 \times 10^{-6}$ cm.), and the 'parallel' crystals T1, T6 and T9 (average $\lambda_0 = 5.21 \times 10^{-6}$ cm.), which, even allowing for possible systematic errors associated with the in-phase correction, seemed significant. Once we were aware of the possibilities of mistaken interpretation due to poor surface conditions, we remeasured T1, T3 and T5 after electrolytic polishing, and it can be seen that no significant difference remains (if anything, there is a suggestion of anisotropy in the opposite sense). It is interesting that although λ_0 for the perpendicular crystals is reduced by polishing there is no such reduction for the parallel crystal, suggesting that perhaps surface cracks occur more easily in perpendicular than in parallel crystals. Such cracks must presumably be on a submicroscopic scale, since all the crystals had the same highly polished appearance even before electrolytic polishing. It was hoped that T10, prepared by the method which proved so successful with mercury, might have a more nearly perfect surface than T3 and T5, but actually there was no significant improvement.

The early measurements on the machined tin cylinder T0 are included in table 2 for completeness; the results for T0 seem to indicate a slightly higher value of λ_0 than do the best crystals; this may be due to slight roughness or to the fact that T_c was not directly measured and had to be estimated from the results themselves, which can easily cause a systematic error of a few per cent in λ_0 .

If λ is indeed slightly anisotropic, and λ_1 and λ_3 are the penetration depths for current flow perpendicular and parallel to the principal crystal axis respectively, the observed penetration depth for a 'parallel' crystal should be λ_1 (since the current flows everywhere perpendicular to the rod axis). For a 'perpendicular' crystal, however, the penetration depth varies round the circumference, and the observed penetration is an average between λ_1 and λ_3 ; as shown elsewhere (Fraser & Shoenberg 1949) the appropriate average for slight anisotropy is the arithmetic mean $\frac{1}{2}(\lambda_1 + \lambda_3)$. More generally, for a crystal whose principal axis is inclined at angle ψ to the rod axis, we should find an effective penetration λ_0 given by

$$\lambda_0 = \frac{1}{2}((\lambda_1 + \lambda_3) + (\lambda_1 - \lambda_3) \cos^2 \psi). \quad (11)$$

Thus any observed difference in λ_0 for parallel and perpendicular single crystals represents only half the difference between λ_1 and λ_3 .

For tin, bearing in mind the rather large uncertainties due to in-phase corrections and assuming that the electrolytically polished specimens have perfect surfaces, the difference between λ_1 and λ_3 is unlikely to exceed 20 %, and there is no definite indication which is the bigger. For mercury, considering only the good specimens

M 13 to 17, figure 11 shows that there is some correlation between λ_0 and the value of ρ at 80° K. As can be seen from (7) and (11), both λ_0 and ρ should vary in the same way with orientation, and it is probably justified to suppose that even for specimens consisting of several crystals (such as M 15) λ_0 and ρ are affected by any preferred orientation in the same way; thus we should expect, if λ is anisotropic, to find a linear relation between λ_0 and ρ . On the basis of the straight line drawn in figure 11, it follows that $\lambda_1 = 4.5 \times 10^{-6}$ cm. and $\lambda_3 = 3.8 \times 10^{-6}$ cm., but this interpretation is speculative, since the random errors of the measurements and possible systematic errors due to surface conditions are of the same order as the observed differences of λ_0 .

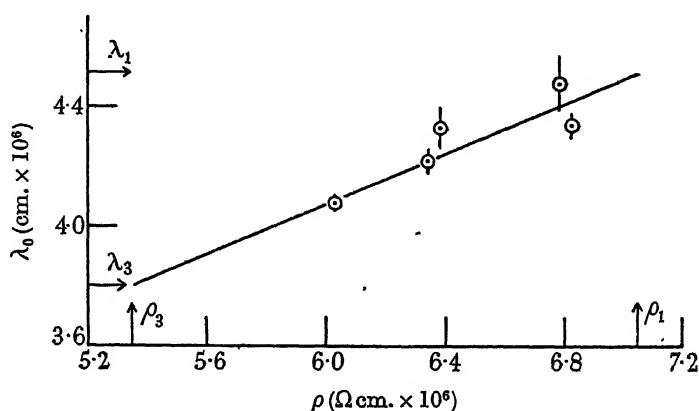


FIGURE 11. Variation of λ with ρ at 80° K for the 'good' mercury specimens M 13 to 17. The vertical lines indicate standard deviations.

As regards absolute mean values, we estimate $\lambda_0 = 5.2 \times 10^{-6}$ cm. for tin and 4.3×10^{-6} cm. for mercury (for a polycrystal we should have $\lambda_0 = \frac{1}{3}(2\lambda_1 + \lambda_3)$). If surface cracks are still present even in the polished tin and the 'good' mercury specimens, these may be overestimates, though the general agreement between specimens makes this improbable. It should be remembered, too, that λ_0 has been calculated on the assumption that (8) applies, which has been reliably confirmed only for mercury; if (8) breaks down at low temperatures, λ_0 will no longer mean the value of λ at 0° K, but merely the parameter which fits the data to equation (9) best.

These estimates of λ_0 for tin and mercury may be compared with some previous estimates. The only previous direct estimate of λ_0 for mercury was that of Désirant & Shoenberg (1948) based on the susceptibility of thin cylinders and gave $\lambda_0 = 7.6 \times 10^{-6}$ cm., which is nearly twice as big as the present estimate. Probably the discrepancy arises because of the troublesome effects described above which occur when mercury is cooled in a container. In the susceptibility method there was no 'tell-tale' such as the 'in-phase' effect to show that anything was wrong, and since a large number of cylinders were always used simultaneously, probably the ratio 7.6 to 4.3 represents some average magnification effect. Whether this magnification is due to a direct increase of perimeter caused by cracked surfaces or to the extra penetration in normal regions persisting below T_c (the earth's magnetic field was not compensated) cannot be decided. Indirect estimates of λ_0 for mercury have

been made by Pippard (1947) from the R.F. resistance and reactance at 1200 Mc./sec., and have also indicated a value of about 7.5×10^{-6} cm. Since it is improbable that our estimate can be too *low*, this discrepancy suggests that Pippard's interpretation of his R.F. results may need revision.

In view of their high estimate of λ_0 , Désirant & Shoenberg's (1948) discussion of Ginsburg's (1945) interpretation of the critical field measurements on thin mercury films by Appleyard, Bristow, London & Misener (1939) requires revision. Combining the present results with the thin film data, Ginsburg's theory gives β (defined as $8\pi(\alpha_n - \alpha_s)/H_c^2$, where α_n and α_s are the surface tensions at a boundary between an insulator and the normal and superconducting phases respectively) as 7×10^{-6} cm. at 0° K, and rising roughly proportionally to λ with rise of temperature. This agrees qualitatively with Ginsburg's own interpretation based on the thin film results alone.

For tin, there have been several previous measurements. The above-mentioned method of Désirant & Shoenberg was applied to one specimen of thin tin cylinders in pyrex capillaries and gave rather higher values of $\Delta\lambda$ than the present experiments, probably partly for the same reasons as in the case of mercury, and partly because the specimen had a spread-out transition and an abnormally high value of T_c . A much more precise determination was that of Pippard (1947), based on the change of reactance of a superconducting resonator at 1200 Mc./sec. when superconductivity is destroyed by a magnetic field. This method could not be applied directly to mercury owing to the variation of the normal resistance with temperature; the interpretation is far more direct than for the other R.F. measurements mentioned above. The temperature variation of $\Delta\lambda$ deduced by Pippard is shown in figure 9,* and it can be seen that below $z = 2$ his results agree well with our T 1 results. The divergences at higher z may be due to inadequacy of the theoretical correction used to allow for the effect of 'normal' electrons, which should, however, be important only for high z (when the proportion of normal electrons grows rapidly). Evidence for the inadequacy of the theory in this region comes from more recent work at 9200 Mc./sec. (Pippard 1948), which shows that the theory, although qualitatively describing the results, does fail in quantitative details.

The only other study of penetration effects in tin is that of Shalnikov & Sharvin (1948), who used a method in which the temperature of an ellipsoid in a steady magnetic field was oscillated, and the resulting e.m.f. in a surrounding coil due to oscillations of penetration depth observed. Absolute values of λ were deduced by integration and extrapolation processes, and show a good linear dependence on z , in agreement with (8). The value of λ_0 is, however, 11.7×10^{-6} cm., which is more than twice our value. Since the method has no 'tell-tale' like our in-phase measurement to give warning of extraneous effects due to normal regions, and since the specimen used by Shalnikov & Sharvin was prepared in a way which might cause roughness of surface on a microscopic scale, it is unlikely that the observed high value is genuine.

* It is probable that the specimen used consisted of a single crystal or a few large crystals, but nothing is known of the orientation.

Our absolute values of λ_0 , which the above discussion suggests are the most reliable available, may be translated into an effective number of 'superconducting' electrons, n_s per cm.³, on the basis of the London & London (1935) relation

$$\lambda = (mc^2/4\pi n_s e^2)^{\frac{1}{2}},$$

where m is the effective mass of such electrons. We can, therefore, deduce the value of the ratio

$$r = \frac{n_s}{n} \frac{m}{m_0},$$

where n is the number of atoms per cm.³ and m_0 is the ordinary electronic mass. Putting in numerical values, we find $r = 0.30$ for tin and $r = 0.35$ for mercury at 0° K. If we suppose $m = m_0$, this means that for both metals only about one electron from every three atoms is effective, or if we suppose all the valence electrons are effective $m = 5.7m_0$ for mercury and $m = 13.3m_0$ for tin.

Dependence of penetration depth on magnetic field

It has been suggested that λ may depend on the strength of an applied magnetic field, i.e. on the strength of the superconducting current (Ginsburg 1947). This can be investigated by looking for a change of M without a corresponding change of M' on application of a steady field H parallel to the specimen axis. To eliminate the change of M' caused by coupling of the circuit of the field-producing solenoid, readings were taken both with and without the battery in circuit, but with the circuit closed in both cases. A good deal of trouble was encountered owing to small variations of solenoid current which caused unsteadiness of the bridge zero, and though the use of a stud switch variable resistance rather than a rheostat to control the current improved the steadiness, it was possible to work with only small fields, i.e. only at temperatures close to T_c .

Figure 12 shows the results obtained with M 16 at 4.075° K. It can be seen that an appreciable apparent change of λ occurs only fairly close to the critical field H_c , and that just there, small changes of λ' (i.e. of M' ; $\Delta\lambda'$ is obtained from $\Delta M'$ by equation (4)) begin to occur. If, in the usual way, a correction for $\Delta\lambda'$ is applied (in this case a factor of 1.04 is appropriate), practically no significant change of M remains up to 90 % of the critical field. Unfortunately, the measurements of $\Delta M'$ were less accurate in this experiment than usual owing to the above-mentioned difficulties and to the necessity of using small measuring current amplitudes; since, moreover, the correction factor is rather uncertain, the existence of a real effect is not quite excluded. All that can be definitely said is that the change of penetration depth with field is certainly less than 2×10^{-6} cm., i.e. less than 10 % of the penetration depth itself up to 80 % of H_c , and there is no convincing evidence of any increase up to 90 % of H_c . Beyond 90 % of H_c the losses rise so steeply that no conclusions can be drawn.

Similar results were found at 4.091 and 4.123° K for the same specimen, and for M 15 at 4.122° K, while for M 3 a steep apparent rise of penetration depth began much earlier (at 60 % H_c) but was accompanied by a steep rise of $\Delta M'$, so that it is

probably not significant. Presumably these rises in $\Delta M'$ as H_c is approached have the same meaning as for approach to T_c in the temperature-variation curves, namely, a growth of the fraction of metal in the normal state.

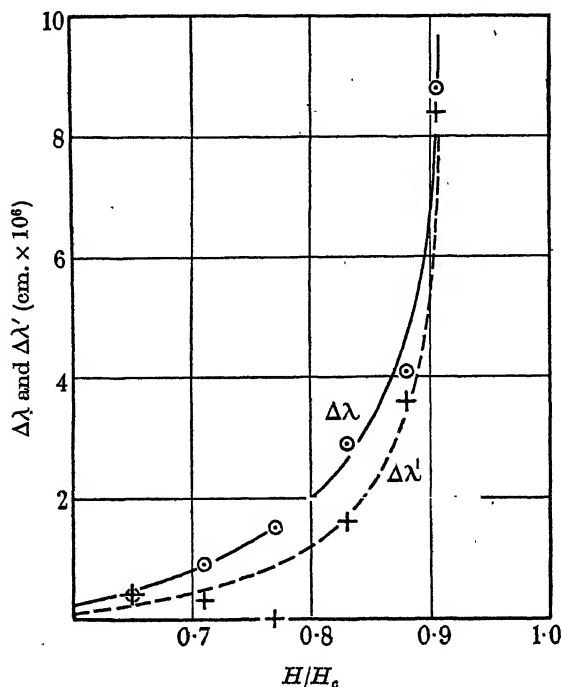


FIGURE 12. Changes of λ and λ' with field for M16 at 4.075°K . The critical field H_c was 15.6 gauss, and the amplitude of the field of the measuring current was 0.5 gauss or $0.03H_c$.

This negative result on the effect of a steady magnetic field in increasing penetration depth suggests that the previous possible indication of such an effect found by Désirant & Shoenberg (1948) must have been due to some other cause.

We wish to thank Professor E. N. da C. Andrade for valuable advice on the preparation of mercury crystals, Mr J. V. Smith of the Crystallographic Laboratory for the X-ray studies of the mercury crystals, Mr J. W. Menter of the Physics and Chemistry of Rubbing Solids Laboratory for electron diffraction studies of tin crystals, and Mr A. B. Pippard for valuable discussion of our results.

REFERENCES

- Andrade, E. N. da C. & Hutchings, P. J. 1935 *Proc. Roy. Soc. A*, **148**, 120.
 Appleyard, E. T. S., Bristow, J. R., London, H. & Misener, A. D. 1939 *Proc. Roy. Soc. A*, **172**, 540.
 Bridgeman, P. W. 1925 *Proc. Amer. Acad. Sci.* **60**, 305.
 Casimir, H. B. G. 1940 *Physica*, **7**, 887.
 Daunt, J. G., Miller, A. R., Pippard, A. B. & Shoenberg, D. 1948 *Phys. Rev.* **74**, 842.
 Désirant, M. & Shoenberg, D. 1948 *Proc. Phys. Soc.* **60**, 413.
 Fraser, A. R. & Shoenberg, D. 1949 *Proc. Camb. Phil. Soc.* (in the Press).

- Ginsburg, V. 1945 *J. Phys. U.S.S.R.* **9**, 305.
 Ginsburg, V. 1947 *J. Phys. U.S.S.R.* **11**, 93.
 Grüneisen, E. & Sckell, O. 1934 *Ann. Phys., Lpz.*, **19**, 387.
 Keesom, W. H. 1932 *Leiden Comm. Suppl.* 71d.
 Laurmann, E. & Shoenberg, D. 1947 *Nature*, **160**, 747.
 London, F. & London, H. 1935 *Proc. Roy. Soc. A*, **149**, 71.
 Pippard, A. B. 1947 *Proc. Roy. Soc. A*, **191**, 399.
 Pippard, A. B. 1948 *Nature*, **162**, 68.
 Sckell, O. 1930 *Ann. Phys., Lpz.*, **6**, 932.
 Shalnikov, A. I. & Sharvin, Y. V. 1948 *Bull. Acad. Sci. U.R.S.S.* (Phys. Series, in Russian), **12**, 195.
 Shoenberg, D. 1940 *Proc. Roy. Soc. A*, **175**, 49.

Electroviscosity. IV. Some extensions of the theory of flow of liquids in narrow channels

BY G. A. H. ELTON AND F. G. HIRSCHLER

Chemistry Department, Battersea Polytechnic

(Communicated by J. Kenyon, F.R.S.—Received 26 March 1949)

In this paper the previous theoretical work on electroviscosity is extended and modified, with special reference to the simplifying assumptions necessary in the derivation of the equations.

1. INTRODUCTION

In previous papers in this series (Elton 1948*a*, 1948*b*, 1949), it has been shown that when an electrical double layer of ions at an interface in an ionic liquid is sheared, the electrical resistance to shear may, in certain circumstances, be an important factor affecting the rate at which flow takes place in the liquid. For instance, when an ionic liquid flows through a narrow channel in a solid, it may exhibit an apparent viscosity equal to several times its normal bulk viscosity. On the basis of various simplifying assumptions, equations were derived to show how the apparent viscosity of a liquid varies under such conditions. For example, it was shown (1948*a*) that when an ionic liquid flows between two similar parallel plates of infinite extent, separated by a distance $2h$, then the apparent viscosity η_a of the liquid is given by

$$\eta_a = \eta + \frac{3\epsilon^2\zeta^2}{32\pi^2\kappa h^3} (h - \lambda + \lambda e^{-h/\lambda}), \quad (1)$$

where η is the normal viscosity of the liquid, ϵ is the dielectric constant, ζ the electrokinetic potential at the solid-liquid interface, κ the specific conductivity of the liquid in the tube, and λ is the Debye-Hückel expression for the effective thickness of the double layer. When $h \gg \lambda$ equation (1) reduces to

$$\eta_a = \eta + \frac{3\epsilon^2\zeta^2}{32\pi^2\kappa h^2}. \quad (2)$$

Two simplifying assumptions were made in the derivation of equation (1), namely, (a) that in calculating the charge distribution in the double layer it was permissible to write $\sinh e\psi/kT = e\psi/kT$, where e is the electronic charge, k the Boltzmann constant, T the absolute temperature, and ψ is the potential at any point in the double layer; and (b) that the potential ψ_m at the axis of the tube was zero.

It will be shown that for a single surface, the first of these approximations, although strictly only a limiting one, nevertheless results in an expression for the potential distribution which is in good agreement with that obtained by use of the full expression for values of ζ as high as 150 mV. The assumption (b) then becomes unnecessary, provided that the agreement persists in the case of two surfaces.

2. THE POTENTIAL DISTRIBUTION AT A SINGLE SURFACE

Considering the ions of the double layer as point charges the Boltzmann distribution equation leads, for uni-univalent electrolytes of ionic concentration n , to the expression

$$\rho_x = -2ne \sinh \frac{e\psi}{kT},$$

where ψ is the potential at a distance x from the solid-liquid interface, and ρ_x is the corresponding excess charge density.

When this is combined with the Poisson equation

$$\frac{d^2\psi}{dx^2} = -\frac{4\pi\rho_x}{\epsilon},$$

the following equation is obtained:

$$\frac{d^2\psi}{dx^2} = \frac{8\pi ne}{\epsilon} \sinh \frac{e\psi}{kT}. \quad (3)$$

If the substitution $e\psi/kT = \sinh e\psi/kT$ is made, this equation becomes

$$\frac{d^2\psi}{dx^2} = \frac{8\pi ne^2}{\epsilon kT} \psi. \quad (4)$$

From these two equations (3) and (4) it follows that for any given value of ψ , $d^2\psi/dx^2$ is greater when calculated from (3) than when calculated from (4), and that $d\psi/dx$ as calculated from (3) is more negative than as calculated from (4). At the surface the condition is that $\psi = \psi_0$, a fixed value. Hence ψ as calculated from (3) drops more rapidly at first than as calculated from (4), but the slope, $d\psi/dx$, obtained from (3) more rapidly levels out than that obtained from (4). Since at $x = 0$ both values of ψ are the same, and these two effects tend to cancel out, the values of ψ never become very far separated.

(i) Solution of equation (3)

Multiplying equation (3) throughout by $d\psi/dx$ and integrating

$$\frac{1}{2} \left(\frac{d\psi}{dx} \right)^2 = \frac{8\pi n kT}{\epsilon} \cosh \frac{e\psi}{kT} + B \quad (5)$$

is obtained, where B is an integration constant. If when $x = \infty$, $\psi = 0$ and $d\psi/dx = 0$, then

$$B = -\frac{8\pi n k T}{\epsilon}$$

and
$$\frac{d\psi}{dx} = \pm 2 \left\{ \frac{8\pi n k T}{\epsilon} \right\}^{\frac{1}{2}} \sinh \frac{e\psi}{2kT}. \quad (6)$$

Therefore
$$\int \frac{d\psi}{\sinh e\psi/2kT} = \pm 2 \left\{ \frac{8\pi n k T}{\epsilon} \right\}^{\frac{1}{2}} \int dx,$$

i.e.
$$\log \tanh \frac{e\psi}{4kT} = \pm \frac{2x}{\lambda} + C. \quad (7)$$

When $x = 0$, $\psi = \psi_0$, hence $C = -\log \tanh e\psi_0/4kT$ and

$$-\frac{x}{\lambda} = \log \frac{\tanh e\psi/4kT}{\tanh e\psi_0/4kT} \quad (8)$$

(the negative root is taken, since for all values of x , $\psi \leq \psi_0$, and

$$\tanh e\psi/4kT \leq \tanh e\psi_0/4kT).$$

Hence
$$e^{-x/\lambda} = \tanh \frac{e\psi}{4kT} / \tanh \frac{e\psi_0}{4kT}. \quad (9)$$

Putting $A = \tanh e\psi_0/4kT$, this may be written

$$e^{e\psi/2kT} = \frac{1 + A e^{-x/\lambda}}{1 - A e^{-x/\lambda}}. \quad (10)$$

Hence
$$\psi = \frac{2kT}{e} \log \frac{1 + A e^{-x/\lambda}}{1 - A e^{-x/\lambda}}. \quad (11)$$

(ii) Solution of equation (4)

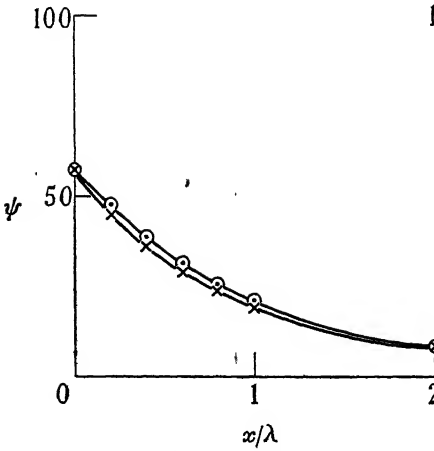
This is a standard equation whose solution is

$$\psi = \alpha e^{x/\lambda} + \beta e^{-x/\lambda},$$

with the boundary conditions $x = 0$, $\psi = \psi_0$; $x = \infty$, $\psi = 0$, this leads to

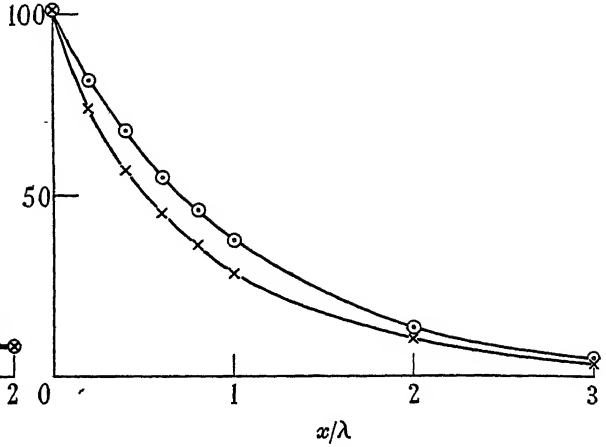
$$\psi = \psi_0 e^{-x/\lambda}. \quad (12)$$

Curves showing the relation between ψ and x are drawn in figures 1 to 4. The values of ψ were calculated from the equations (11) and (12) respectively, using different values of ψ_0 , the potential at the surface ($= \zeta$). These curves show that the values of ψ as calculated from equation (12) are in good agreement with those calculated from equation (11) for values of ζ as high as 150 mV. It may therefore be inferred that equation (4), and hence equation (12) also, give a good approximation to the true equations of the potential distribution.



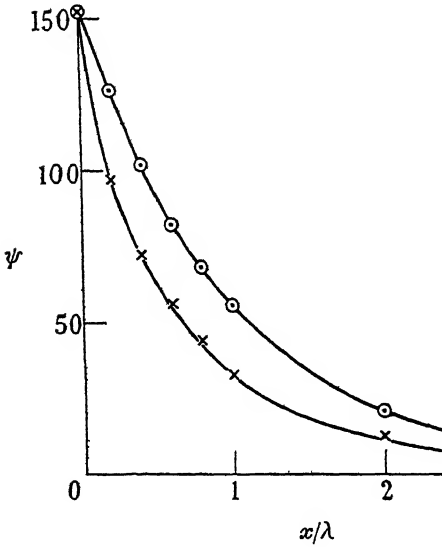
○ refers to equation (12)

FIGURE 1. In equation (11) $A = 0.5$, hence $\psi_0 = 56.8$ mV.



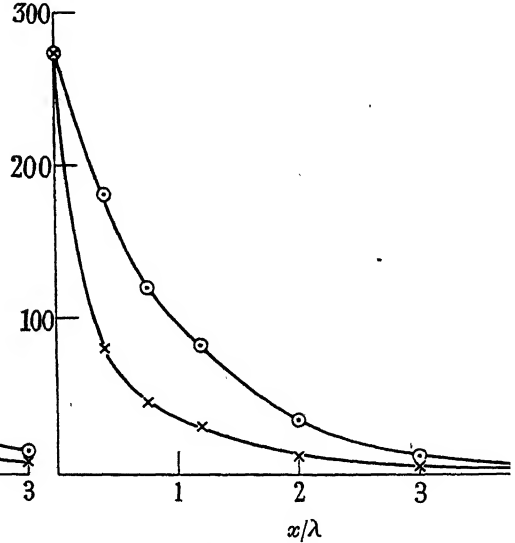
× refers to equation (11)

FIGURE 2. In equation (11) $A = 0.75$, hence $\psi_0 = 100.7$ mV.



○ refers to equation (12)

FIGURE 3. In equation (11) $A = 0.9$, hence $\psi_0 = 152.5$ mV.



× refers to equation (11)

FIGURE 4. In equation (11) $A = 0.99$, hence $\psi_0 = 274$ mV.

3. POTENTIAL DISTRIBUTION BETWEEN TWO INFINITE PARALLEL PLATES DISTANCE $2h$ APART

The full equation (3) cannot be solved in general terms in this case. The boundary conditions that $\psi = \psi_0$ when $x = 0$; $\psi = \psi_m$ and $d\psi/dx = 0$ when $x = h$; and $\psi = \psi_0$ when $x = 2h$, change equation (6) to

$$\frac{d\psi}{dx} = \left\{ \frac{16\pi n k T}{\epsilon} (\cosh y - \cosh u) \right\}^{\frac{1}{2}}, \quad (13)$$

where $y = e\psi/kT$, and $u = e\psi_m/kT$. This equation, on integration, gives an elliptic integral involving ψ . ψ may then be evaluated for various values of x . The function is of the form

$$F(\alpha', \phi) = \frac{x}{2\lambda\sqrt{(\sin \alpha')}} + C, \quad (14)$$

where $\log \sin \alpha' = e\psi_m/kT$ and $\log \sin \phi = e(\psi_m - \psi)/kT$, and C is an integration constant.

The problem may be solved if the approximation $\sinh e\psi/kT = e\psi/kT$ is made, and equation (3) reduced to equation (4). It has been shown in § 2 that this approximation leads to good results for a single surface. Its use for the potential distribution between a pair of parallel plates is therefore not yet justified, but it may be concluded that for plates not extremely close together the equations obtained represent the potential distribution quite well. The present theory is in any case invalid for plates in very close proximity (where h/λ is less than 1), for then the number of ions concerned is generally no longer statistical, especially when the effect of the ionic atmosphere in gathering together of the ions is remembered. These last considerations limit the use of the theory to plates at a moderate distance apart, at low concentration because of the high value of λ , and at higher concentrations because of effect of non-statistical numbers of ions between the plates (λ and hence also h being small).

Under these limitations equation (4) will now be solved for the case of a pair of parallel plates of infinite extent.

Multiplying the equation throughout by $d\psi/dx$ and integrating

$$\frac{1}{2} \left(\frac{d\psi}{dx} \right)^2 = \frac{\psi^2}{2\lambda^2} + \text{constant}$$

is obtained. Putting in the condition $d\psi/dx = 0$, $\psi = \psi_m$ at $x = h$, taking the square root, and integrating again

$$-\frac{x}{\lambda} = \log \frac{[\psi + \sqrt{(\psi^2 - \psi_m^2)}]}{[\psi_0 + \sqrt{(\psi_0^2 - \psi_m^2)}]} \quad (15)$$

is obtained. Solving this equation for ψ , we obtain, for values of x between 0 and h ,

$$\psi = \frac{1}{2}A' e^{-x/\lambda} + \frac{\psi_m^2}{A'} e^{x/\lambda}, \quad (16)$$

where

$$A' = \psi_0 + \sqrt{(\psi_0^2 - \psi_m^2)}.$$

On putting the condition $\psi = \psi_m$ when $x = h$ in equation (15), the following relation between ψ_m and ψ_0 is obtained:

$$\psi_m = \frac{\psi_0}{\cosh h/\lambda}. \quad (17)$$

4. DERIVATION OF THE EQUATION OF FLOW BETWEEN PARALLEL PLATES

As previously shown (1948a), the quantity of liquid Q flowing per second through a section of the system of unit length and breadth under a pressure gradient P is

$$Q = \int_0^{2h} v dx = \frac{1}{\eta} \int_0^{2h} \left[\frac{Px}{2} (2h - x) + \frac{eE}{4\pi} (\psi - \psi_0) \right] dx, \quad (18)$$

where v is the rate of flow at distance x from one plate, and E is the streaming potential gradient set up in the plane of flow. Substituting for ψ from equation (16)

$$Q = \frac{2}{\eta} \int_0^h \left[\frac{Px}{2} (2h-x) - \frac{\epsilon E}{4\pi} \left(\psi_0 - \frac{1}{2} A' e^{-x/\lambda} - \frac{\psi_m^2}{2A'} e^{x/\lambda} \right) \right] dx, \quad (19)$$

since by symmetry
$$\int_0^{2h} v dx = 2 \int_0^h v dx.$$

Equation (19) on integration gives

$$Q = \frac{2Ph^3}{3\eta} - \frac{\epsilon E}{4\pi\eta} \left[2\psi_0 x + \lambda A' e^{-x/\lambda} - \frac{\lambda\psi_m^2}{A'} e^{x/\lambda} \right]_0^h. \quad (20)$$

Let the reduced rate of flow result in an apparent viscosity η_a , where

$$Q = \frac{2Ph^3}{3\eta_a}, \quad E = \frac{\zeta\epsilon P}{8\pi\eta_a\kappa} \quad \text{and} \quad \psi_0 = \zeta.$$

Then
$$\eta_a = \eta + \frac{3\epsilon^2\zeta}{32\pi^2\kappa h^3} \left[\zeta h - \frac{\lambda A' (1 - e^{-h/\lambda})}{2} - \frac{\lambda\psi_m^2 (e^{h/\lambda} - 1)}{2A'} \right], \quad (21)$$

where $A' = \zeta + \sqrt{(\zeta^2 - \psi_m^2)}$. From equation (21) we can calculate values of η_a when ψ_m is not zero. The required values of ψ_m may be obtained from equation (17), or they may be calculated from the full equation by means of the solution (14). Values of ψ_m for various values of h/λ and ψ_0 obtained in this way, are given by Verwey & Overbeek (1948). When ψ_m is substituted from equation (17) equation (21) is reduced to

$$\eta_a = \eta + \frac{3\epsilon^2\zeta^2}{32\pi^2\kappa h^3} \left(h - \tanh \frac{h}{\lambda} \right). \quad (22)$$

It should be noted here that the term ζ^2 appearing in equation (22) is not a consequence of the approximation $e\psi/kT = \sinh e\psi/kT$, but one factor ζ comes from the expression for the streaming potential E , and the other from the integration $\int \psi dx$. Hence the error introduced by the approximation appears in the final equation only in the magnitude of the error in the integral $\int \psi dx$. From the figures 1 to 4 it will be seen that for values of ζ less than 150 mV the areas under the two curves differ only very little. In table 1 values of F' the ratio of $\eta_a - \eta$ calculated from equation (22) to those calculated from equation (1) are given for different systems. For moderate values of h/λ the agreement is good.

5. DERIVATION OF FLOW EQUATION BETWEEN PARALLEL PLATES SEPARATED BY A LARGE DISTANCE

For plates separated by a distance $2h$, where h/λ is greater than 10, ψ_m , the potential at the centre of the tube, is less than $\frac{1}{1000}$ th of ψ_0 , the potential at the surface. Consequently, the assumption that the double layers at the two surfaces do not interfere is valid. The electroviscous effect may then be calculated without making the approximation $\sinh e\psi/kT = e\psi/kT$.

TABLE I

h (cm.)	electrolyte normality	h/λ	F'
1×10^{-5}	1×10^{-6}	0.32	0.14
	1×10^{-5}	1.03	0.70
	1×10^{-4}	3.24	0.98
	4×10^{-4}	6.49	1.00
3×10^{-5}	1×10^{-6}	0.97	0.65
	1×10^{-5}	3.08	1.00
	1×10^{-4}	9.73	1.00
5×10^{-5}	1×10^{-6}	1.62	0.95
	1×10^{-5}	5.13	1.00
	1×10^{-4}	16.2	1.00

Since there is no interference the potential distribution at both the surfaces is the same, and the same as that for a single double layer, hence the distribution is given by equation (11). Substituting for ψ in equation (18) from equation (11), expanding the logarithm and integrating, the expression

$$Q = \frac{2P\hbar^3}{3\eta} - \frac{eE\psi_0\hbar}{2\pi\eta} + \frac{2ekT\lambda E}{e\pi\eta} \left[A(1 - e^{-h/\lambda}) + \frac{A^3}{3^2}(1 - e^{-3h/\lambda}) + \frac{A^5}{5^2}(1 - e^{-5h/\lambda}) + \dots \right]$$

is obtained. Substituting for Q , E and ψ_0 as before and rearranging

$$\eta_a = \eta + \frac{3e^2\zeta^2}{32\pi^2\kappa\hbar^2} - \frac{3e^2\zeta kT\lambda}{8e\pi^2\kappa\hbar^3} \left[A(1 - e^{-h/\lambda}) + \frac{A^3}{3^2}(1 - e^{-3h/\lambda}) + \frac{A^5}{5^2}(1 - e^{-5h/\lambda}) + \dots \right]. \quad (22a)$$

6. FLOW IN A CYLINDRICAL TUBE. $\psi_m = 0$, $\sinh e\psi/kT = e\psi/kT$

This case has been examined in a previous paper (1948a). It is proposed here to give a full solution of the problem, and to demonstrate that the equation reduces in the limit to that obtained previously by an approximate method.

(i) Potential distribution in a cylindrical tube

The Poisson-Boltzmann distribution equation for a cylindrical tube is, when $\sinh e\psi/kT = e\psi/kT$,

$$\frac{1}{r} \frac{d}{dr} \left(r \frac{d\psi}{dr} \right) = \frac{\psi}{\lambda^2}, \quad (23)$$

where ψ is the potential at a point distant r from the axis of the tube. In part I of this series (1948a), equation (9) is not a correct solution to this differential equation. This solution should be

$$\psi = MI_0\left(\frac{r}{\lambda}\right) + NK_0\left(\frac{r}{\lambda}\right), \quad (24)$$

where $I_0(r/\lambda)$ is a Bessel function of imaginary argument of the first kind and zero order and $K_0(r/\lambda)$ is a Bessel function of the second kind and zero order. M and N are constants of integration.

For a cylindrical tube of radius a , when $\psi = 0$ for $r = 0$ and $\psi = \psi_0$ for $r = a$, $N = 0$ and

$$\psi = \psi_0 \frac{I_0(r/\lambda)}{I_0(a/\lambda)}. \quad (25)$$

(ii) *Derivation of the flow equation*

The quantity of liquid passing through the tube in unit time is given by

$$Q = \int_0^a 2\pi r v dr, \quad (26)$$

where v , the rate of flow of liquid at distance r from the axis, is given by

$$v = \frac{P(a^2 - r^2)}{4\eta} - \frac{\epsilon E}{4\pi} (\psi_0 - \psi). \quad (27)$$

Substituting from equations (25) and (27) in equation (26) and integrating, we obtain

$$Q = \frac{\pi P a^4}{8\eta} - \frac{\epsilon E \psi_0}{2\eta} \int_0^a r \left(1 - \frac{I_0(r/\lambda)}{I_0(a/\lambda)} \right) dr. \quad (28)$$

Integrating the second term by parts we obtain

$$Q = \frac{\pi P a^4}{8\eta} - \frac{\epsilon E \psi_0}{2\eta} \left[\frac{a^2}{2} - \frac{1}{I_0(a/\lambda)} \int_0^a r I_0(r/\lambda) dr \right],$$

which on use of the properties of Bessel functions reduces to

$$Q = \frac{\pi P a^4}{8\eta} - \frac{\epsilon E \psi_0 a^2}{4\eta} \left[1 - \frac{2\lambda}{a} \frac{I_1(a/\lambda)}{I_0(a/\lambda)} \right]. \quad (29)$$

If the reduced rate of flow results in an apparent viscosity η_a we obtain, on putting

$$Q = \frac{\pi P a^4}{8\eta_a}, \quad E = \frac{\zeta e P}{4\pi \eta_a \kappa} \quad \text{and} \quad \psi_0 = \zeta, \quad (30)$$

$$\eta_a = \eta + \frac{\epsilon^2 \zeta^2}{2\pi^2 \kappa a^2} \left[1 - \frac{2\lambda}{a} \frac{I_1(a/\lambda)}{I_0(a/\lambda)} \right].$$

For values of a/λ greater than 9, the first two terms of the asymptotic expansions of the Bessel functions give accurate values. Hence

$$\frac{2\lambda}{a} \frac{I_1(a/\lambda)}{I_0(a/\lambda)} = \frac{2\lambda}{a} \frac{(1 - 3\lambda/8a)}{(1 + \lambda/8a)}.$$

Using the binomial theorem this reduces to

$$\frac{2\lambda}{a} \frac{I_1(a/\lambda)}{I_0(a/\lambda)} = \frac{2\lambda}{a} \left(1 - \frac{3\lambda}{8a} \right) \left(1 - \frac{\lambda}{8a} \right) \simeq \frac{2\lambda}{a} \left(1 - \frac{\lambda}{2a} \right).$$

Hence

$$1 - \frac{2\lambda}{a} \frac{I_1(a/\lambda)}{I_0(a/\lambda)} \simeq 1 - \frac{2\lambda}{a} + \frac{\lambda^2}{a^2} = \left(1 - \frac{\lambda}{a} \right)^2$$

and

$$\eta_a = \eta + \frac{\epsilon^2 \zeta^2}{2\pi^2 \kappa a^2} \left(1 - \frac{\lambda}{a} \right)^2 \simeq \eta + \frac{\epsilon^2 \zeta^2}{2\pi^2 \kappa a^2}. \quad (31)$$

In this equation, as in equation (22) of § 4, the term ζ^2 does not arise from the approximation $\sinh e\psi/kT = e\psi/kT$, but one factor comes from the expression for E , the streaming potential gradient, and the other from the integral $\int \psi dr$. As shown in § 6 the error involved is therefore of the order of the error in the integral $\int \psi dr$.

7. CONCLUSION

It has been shown in this paper that the Debye-Hückel approximation, $\sinh e\psi/kT = e\psi/kT$, leads to potential distribution equations at single flat surfaces which are in good agreement with those calculated from the full expression, for values of the electrokinetic potential up to 150 mV. It has been deduced that the electroviscous effect between parallel plates calculated from these equations is not greatly in error for all plate separations for which viscosity can be treated as a statistical phenomenon. An expression for the electroviscous effect in a cylindrical capillary tube was obtained, making the assumption that the potential at the axis is zero. By analogy with the case of parallel plates this assumption will be good if the ratio radius of capillary to double-layer thickness is greater than 10. The expression is useful for all capillaries likely to be encountered in practice. For example, in a capillary of radius 1μ , η_a may still be appreciably larger than η . The error here will be larger than for the case of a corresponding parallel-plate system, since the potential at the axis will be higher than in the case of parallel plates.

The authors wish to express their thanks to Dr K. L. Sutherland and Dr W. E. Ewers for their interest and assistance in this work.

REFERENCES

- Elton, G. A. H. 1948*a* *Proc. Roy. Soc. A*, **194**, 259.
 Elton, G. A. H. 1948*b* *Proc. Roy. Soc. A*, **194**, 275.
 Elton, G. A. H. 1949 *Proc. Roy. Soc. A*, **197**, 568.
 Verwey, E. J. W. & Overbeek, J. Th. G. 1948 *Theory of the stability of lyophobic colloids*.
 London: Elsevier Publishing Company Inc.

INDEX TO VOLUME 198 (A)

- Anisotropic metals, plane plastic strain (Hill), 428.
- Auluck, F. C. & Kothari, D. S. A note on the Riesz method and the method of residues, 170.
- Autoxidation of tetralin (Bamford & Dewar), 252.
- Baker, J. F., Horne, M. R. & Roderick, J. W. The behaviour of continuous stanchions, 493.
- Bamford, C. H. & Dewar, M. J. S. The autoxidation of tetralin, 252.
- Barber, N. F. The behaviour of waves on tidal streams, 81.
- Behaviour of continuous stanchions (Baker, Horne & Roderick), 493.
- Behaviour of waves on tidal streams (Barber), 81.
- Bell, R. P., Gelles, E. & Möller, E. Kinetics of the base-catalyzed halogenation of some ketones and esters, 308.
- Bennett, J. G., Brown, R. L. & Thring, M. W. Unified field theory in a curvature-free five-dimensional manifold, 39.
- Benzylamine C-N bond, dissociation energy (Szwarc), 285.
- Birks, J. & Bradley, R. S. The rate of evaporation of droplets. II. The influence of changes of temperature and of the surrounding gas on the rate of evaporation of drops of di-*n*-butyl phthalate, 226.
- Birth and growth of explosion in liquids and solids initiated by impact and friction (Bowden & Gurton), 350.
- Bleaney, B., Penrose, R. P. & Plumpton, B. I. Paramagnetic resonance in the copper Tutton salts, 406.
- Bowden, F. P. & Gurton, O. A. Birth and growth of explosion in liquids and solids initiated by impact and friction, 350.
- Bowden, F. P. & Gurton, O. A. Initiation of solid explosives by impact and friction: the influence of grit, 337.
- Bradley, R. S. *See* Birks & Bradley.
- Bradley, R. S. & Shellard, A. D. The rate of evaporation of droplets. III. Vapour pressures and rates of evaporation of straight-chain paraffin hydrocarbons, 239.
- Brown, R. L. *See* Bennett, Brown & Thring.
- Catalytic hydrogenation of methyl elaeostearate, and of mixtures of elaeostearic with other polyethenoid long-chain esters (Hilditch & Pathak), 323.
- Copper Tutton salts, paramagnetic resonance (Bleaney, Penrose & Plumpton), 406.
- Corner, J. The effect of diffusion of the main reactants on flame speeds in gases, 388.
- Dewar, M. J. S. *See* Bamford & Dewar.
- Diffraction of blast. I (Lighthill), 454.
- Diffusion in gases and liquids. I, II (Yang), 94, 471.
- Dissociation energy of the C-N bond in benzylamine (Szwarc), 285.
- Dissociation energy of the N-H bond in hydrazine (Szwarc), 267.
- Dixon-Lewis, G. Studies in polymerization. V. The polymerization of vinyl acetate, 510.
- Eddy diffusion of water vapour and heat near the ground (Pasquill), 116.
- Eden, R. J. Heisenberg's *S* matrix for a system of many particles, 540.
- Effect of diffusion of the main reactants on flame speeds in gases (Corner), 388.
- Electroviscosity. IV. Some extensions of the theory of flow of liquids in narrow channels (Elton & Hirschler), 581.

- Elton, G. A. H. & Hirschler, F. G. Electroviscosity. IV. Some extensions of the theory of flow of liquids in narrow channels, 581.
- Explosion in liquids and solids (Bowden & Gurton), 337.
- Explosion in liquids and solids (Yoffe), 373.
- Flame speeds in gases (Corner), 388.
- Frank, F. C. & Merwe, J. H. van der. One-dimensional dislocations. I. Static theory, 205.
- Frank, F. C. & Merwe, J. H. van der. One-dimensional dislocations. II. Misfitting monolayers and orientated overgrowth, 216.
- Frost, R. A note on polar air-mass modification, 27.
- Gelles, E. *See* Bell, Gelles & Möller.
- Gurton, O. A. *See* Bowden & Gurton.
- Harington, Sir Charles. The work of the National Institute for Medical Research, 293.
- Heats of formation of free CN and free CH_2 , and the relationship between $D(\text{CO})$, $D(\text{CN})$ and $D(\text{N}_2)$ (Long), 62.
- Heisenberg's S matrix for a system of many particles (Eden), 540.
- Helium II, film phenomena (Temperley), 438.
- Hilditch, T. P. & Pathak, S. P. The catalytic hydrogenation of methyl elaeostearate, and of mixtures of elaeostearic with other polyethenoid long-chain esters, 323.
- Hill, R. The theory of plane plastic strain for anisotropic metals, 428.
- Hirschler, F. G. *See* Elton & Hirschler.
- Horne, M. R. *See* Baker, Horne & Roderick.
- Hot-wire investigation of the wake behind cylinders at low Reynolds numbers (Kovácsnay), 174.
- Hydrazine N-N bond, dissociation energy (Szwarc), 267.
- Ignition of solid explosive media by hot wires (Jones), 523.
- Influence of entrapped gas on initiation of explosion in liquids and solids (Yoffe), 373.
- Initiation of solid explosives by impact and friction: the influence of grit (Bowden & Gurton), 337.
- Jones, E. The ignition of solid explosive media by hot wires, 523.
- Jones, Sir Harold Spencer. The Royal Greenwich Observatory, 141.
- Kaminski, A. & McBain, J. W. Spontaneous emulsification of pure xylene in an aqueous solution through mere adsorption of a detergent in the interface, 447.
- Ketones and esters, halogenation (Bell, Gelles & Möller), 308.
- Kinetic theory of diffusion in gases and liquids. I. Diffusion and the Brownian motion (Yang), 94.
- Kinetic theory of diffusion in gases and liquids. II. General kinetic theory of liquids mixtures (Yang), 471.
- Kinetics of the base-catalyzed halogenation of some ketones and esters (Bell, Gelles & Möller), 308.
- Kothari, D. S. *See* Auluck & Kothari.
- Kovácsnay, L. S. G. Hot-wire investigation of the wake behind cylinders at low Reynolds numbers, 174.
- Laurmann, E. & Shoenberg, D. Penetration of magnetic field into superconductors. II. Measurements by the Casimir method, 560.
- Lennard-Jones, Sir John. The molecular orbital theory of chemical valency. I. The determination of molecular orbitals, 1.

- Lennard-Jones, Sir John. The molecular orbital theory of chemical valency. II. Equivalent orbitals in molecules of known symmetry, 14.
- Lighthill, M. J. The diffraction of blast. I, 454.
- Liquids, flow (Elton & Hirschler), 581.
- Long, L. H. The heats of formation of free CN and free CH_2 , and the relationship between $D(\text{CO})$, $D(\text{CN})$ and $D(\text{N}_2)$, 62.
- McBain, J. W. *See* Kaminski & McBain.
- Merwe, J. H. van der. *See* Frank & van der Merwe.
- Molecular orbital theory of chemical valency. I. The determination of molecular orbitals (Lennard-Jones), 1.
- Molecular orbital theory of chemical valency. II. Equivalent orbitals in molecules of known symmetry (Lennard-Jones), 14.
- Möller, E. *See* Bell, Gelles & Möller.
- National Institute for Medical Research (Harington), 293.
- Note on polar air-mass modification (Frost), 27.
- Note on the Riesz method and the method of residues (Auluck & Kothari), 170.
- Nye, J. F. Plastic deformation of silver chloride. I. Internal stresses and the glide mechanism, 190.
- One-dimensional dislocations. I. Static theory (Frank & van der Merwe), 205.
- One-dimensional dislocations. II. Misfitting monolayers and oriented overgrowth (Frank & van der Merwe), 216.
- Paramagnetic resonance in the copper Tutton salts (Bleaney, Penrose & Plumpton), 406.
- Pasquill, F. Eddy diffusion of water vapour and heat near the ground, 116.
- Pathak, S. P. *See* Hilditch & Pathak.
- Penetration of magnetic field into superconductors. II. Measurements by the Casimir method (Laurmann & Shoenberg), 560.
- Penrose, R. P. *See* Bleaney, Penrose & Plumpton.
- Plastic deformation of silver chloride. I. Internal stresses and the glide mechanism (Nye), 190.
- Plumpton, B. I. *See* Bleaney, Penrose & Plumpton.
- Polar air-mass modification (Frost), 27.
- Polymerization (Dixon-Lewis), 510.
- Rate of evaporation of droplets. II. The influence of changes of temperature and of the surrounding gas on the rate of evaporation of drops of di-*n*-butyl phthalate (Birks & Bradley), 226.
- Rate of evaporation of droplets. III. Vapour pressures and rates of evaporation of straight-chain paraffin hydrocarbons (Bradley & Shellard), 239.
- Riesz method and method of residues (Auluck & Kothari), 170.
- Roderick, J. W. *See* Baker, Horne & Roderick.
- Royal Greenwich Observatory (Jones), 141.
- Shellard, A. D. *See* Bradley & Shellard.
- Shoenberg, D. *See* Laurmann & Shoenberg.
- Silver chloride, plastic deformation (Nye), 190.
- Solid explosives, ignition (Jones), 523.
- Solid explosives, initiation (Bowden & Gurton), 337.
- Spontaneous emulsification of pure xylene in an aqueous solution through mere adsorption of a detergent in the interface (Kaminski & McBain), 447.

- Stanchions, behaviour (Baker, Horne & Roderick), 493.
- Studies in polymerization. V. The polymerization of vinyl acetate (Dixon-Lewis), 510.
- Superconductors, penetration of magnetic field (Laurmann & Shoenberg), 560.
- Szwarc, M. The dissociation energy of the C-N bond in benzylamine, 285.
- Szwarc, M. The dissociation energy of the N-N bond in hydrazine, 267.
- Temperley, H. N. V. A theory of the film phenomena of liquid helium II, 438.
- Tetralin, autoxidation (Bamford & Dewar), 252.
- Theory of plane plastic strain for anisotropic metals (Hill), 428.
- Theory of the film phenomena of liquid helium II (Temperley), 438.
- Thring, M. W. *See* Bennett, Brown & Thring.
- Tidal streams, waves on (Barber), 81.
- Unified field theory in a curvature-free five-dimensional manifold (Bennett, Brown & Thring), 39.
- Wake behind cylinders, hot-wire investigation (Kovácsnay), 174.
- Water vapour and heat, eddy diffusion of (Pasquill), 116.
- Work of the National Institute for Medical Research (Harington), 293.
- Xylene, spontaneous emulsification (Kaminski & McBain), 447.
- Yang, L. M. Kinetic theory of diffusion in gases and liquids. I. Diffusion and the Brownian motion, 94.
- Yang, L. M. Kinetic theory of diffusion in gases and liquids. II. General kinetic theory of liquids mixtures, 471.
- Yoffe, A. Influence of entrapped gas on initiation of explosion in liquids and solids, 373.

I.A.R.I. 75

INDIAN AGRICULTURAL RESEARCH
INSTITUTE LIBRARY, NEW DELHI.

[illegible]

GIPNLK—H-40 I.A.R I.—29-4-55—15,000